Operations Research Proceedings

Dennis Huisman
Ilse Louwerse
Albert P. M. Wagelmans *Editors*

Operations Research Proceedings 2013

Selected Papers of the International Conference on Operations Research, OR2013, organized by the German Operations Research Society (GOR), the Dutch Society of Operations Research (NGB) and Erasmus University Rotterdam, September 3–6, 2013

GOR
Gesellschaft für Operations Research e.V.
Dutch Society of Operations Research
Springer

Operations Research Proceedings

GOR (Gesellschaft für Operations Research e.V.)

For further volumes:
http://www.springer.com/series/722

Dennis Huisman · Ilse Louwerse
Albert P. M. Wagelmans
Editors

Operations Research Proceedings 2013

Selected Papers of the International Conference on Operations Research, OR2013, organized by the German Operations Research Society (GOR), the Dutch Society of Operations Research (NGB) and Erasmus University Rotterdam, September 3–6, 2013

Springer

Editors
Dennis Huisman
Ilse Louwerse
Albert P. M. Wagelmans
Econometric Institute, Erasmus School of Economics
Erasmus University Rotterdam
Rotterdam
The Netherlands

ISSN 0721-5924
ISBN 978-3-319-07000-1 ISBN 978-3-319-07001-8 (eBook)
DOI 10.1007/978-3-319-07001-8
Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

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

Preface

This volume contains 68 short refereed papers presented at the International Conference on Operations Research, OR2013, which took place in Rotterdam from September 3 to 6, 2013. This conference was jointly organized by the German Operations Research Society (GOR), the Dutch Operations Research Society (NGB), and Erasmus University Rotterdam. The latter was celebrating its centennial in 2013. The conference attracted about 500 participants from all over the world. In total, there were 363 talks (3 plenary, 12 semi-plenary, and 348 contributed talks).

The chapters in this volume are ranked in alphabetical order of the first author. The volume contains chapters from all streams of the conference. The list of streams and stream chairs can be found below:

- Applied Probability and Stochastic Programming, Forecasting (Prof. Richard Boucherie)
- Continuous Optimization (Prof. Mirjam Dür)
- Decision Analysis and Multiple Criteria Decision Making (Prof. Martin Geiger)
- Discrete and Combinatorial Optimization, Graphs and Networks (Prof. Stan van Hoesel)
- Energy and Environment (Prof. Rüdiger Schultz)
- Financial Modeling, Banking and Insurance (Prof. Jörn Sass)
- Game Theory and Experimental Economics (Prof. Stefan Pickl)
- Health Care Management (Dr. Erwin Hans)
- Information Systems, Neural Nets and Fuzzy Systems (Dr. Hans-Georg Zimmermann)
- Managerial Accounting (Prof. Katja Schimmelpfeng)
- Maritime Logistics (Prof. Rob Zuidwijk)
- Production and Operations Management (Prof. Ruud Teunter)
- Revenue Management and Pricing (Prof. Robert Klein)
- Scheduling and Project Management (Prof. Erwin Pesch and Prof. Florian Jaehn)
- Simulation and System Dynamics (Prof. Henk Akkermans)

- Software Applications and Modelling Systems (Prof. Joaquim Gromicho)
- Supply Chain Management, Logistics and Inventory (Prof. Dolores Romero-Morales)
- Traffic and Transportation (Prof. Leena Suhl)

Moreover, the volume contains papers from different award winners:

- Timo Berthold (Young Participant with Most Academic Impact)
- Kirsten Hoffmann (GOR Master Award)
- Max Klimm (GOR Ph.D. Award)
- Jenny Nossack (GOR Ph.D. Award)
- Alena Otto (GOR Ph.D. Award)
- Christian Raack (GOR Ph.D. Award)
- Roman Rischke (GOR Master Award)
- Lara Wiesche (GOR Master Award)

We would like to congratulate all these award winners and Mariel Lavieri, the winner of the Young Participant with Most Practical Impact award.

Finally, we would like to thank everyone who contributed to the organization of this conference. In particular, we would like to mention the members of the program and organizing committee, the stream chairs, the support staff and, last but not least, our sponsors.

Rotterdam, March 2014

Dennis Huisman
Ilse Louwerse
Albert P. M. Wagelmans

Contents

Computing an Upper Bound for the Longest Edge in an Optimal TSP-Solution

Hans Achatz and Peter Kleinschmidt

Abstract A solution of the traveling salesman problem (TSP) with n nodes consists of n edges which form a shortest tour. In our approach we compute an upper bound u for the longest edge which could be in an optimal solution. This means that every edge longer than this bound cannot be in an optimal solution. The quantity u can be computed in polynomial time. We have applied our approach to different problems of the TSPLIB (library of sample instances for the TSP). Our bound does not necessarily improve the fastest TSP-algorithms. However, the reduction of the number of edges might be useful for certain instances.

1 Introduction

The traveling salesman problem (TSP) is one of the most studied problems in combinatorial optimization and has got applications in many different areas. The TSP consists of finding a shortest tour in a complete graph whose edges (i,j) have cost (distance) c_{ij}. A comprehensive treatment of the traveling salesman problem can be found in [3].

In this paper we do not assume that the cost matrix is symmetric. However, our figures will refer to symmetric instances. We consider a dual relaxation of the original problem—the assignment problem A based on the same cost matrix. The result is a dual relaxation, possibly with subtours, as shown in Fig. 1. This problem can also be solved by any code for the assignment problem, e.g. [1].

H. Achatz (✉) · P. Kleinschmidt
University of Passau, Innstr. 39, 94036 Passau, Germany
e-mail: hans.achatz@uni-passau.de

P. Kleinschmidt
e-mail: pkleinschmidt@t-online.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_1,

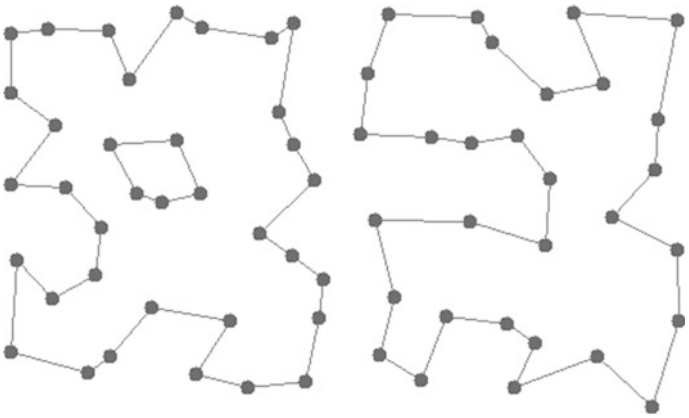


Fig. 1 Dual relaxation with subtours

$$A: \qquad \min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{1}$$

s.t.

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad 1 \leq j \leq n \tag{2}$$

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad 1 \leq i \leq n \tag{3}$$

$$0 \leq x_{ij} \leq 1 \qquad 1 \leq i, j \leq n \tag{4}$$

This optimal solution can be transformed into a tour as shown in the figure below. The value of the objective function of the solution in this example is 2,744 and it is an upper bound for the optimal solution (Fig. 2). By using the Lin-Kernighan heuristic [4] we can obtain an even better upper bound 2,726. The optimal value of the dual heuristic is 2,426. Hence, the length of an optimal tour is between these two values.

2 Computing an Upper Bound

In this paper we introduce a new relaxation A' of the TSP. Due to inequality 6 and 7 every node must have at least one adjacent edge and at most two adjacent edges. Equation 8 assures that there are exactly $n - 1$ edges.

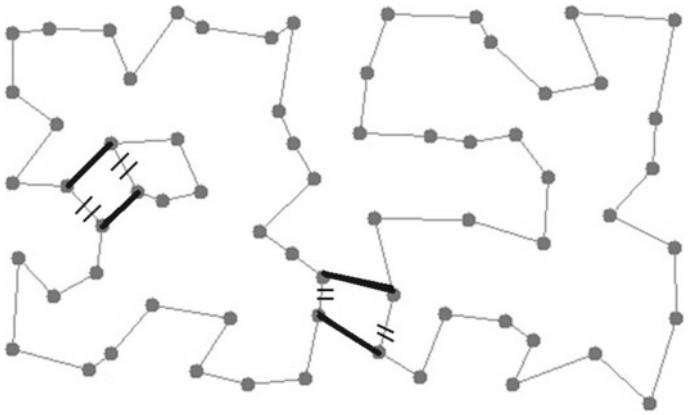


Fig. 2 Transformation into a primal feasible solution

$$A': \qquad \min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{5}$$

$$s.t.$$

$$\sum_{i=1}^{n} x_{ij} + \sum_{i=1}^{n} x_{ji} \leq 2 \qquad 1 \leq j \leq n \tag{6}$$

$$\sum_{i=1}^{n} x_{ij} + \sum_{i=1}^{n} x_{ji} \geq 1 \qquad 1 \leq j \leq n \tag{7}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} = n - 1 \tag{8}$$

$$M, \text{ a set of valid TSP constraints} \tag{9}$$

$$0 \leq x_{ij} \leq 1 \qquad 1 \leq i, j \leq n \tag{10}$$

The set M may consist of some valid TSP constraints which do not contradict constraint 8. For example, M could be chosen as a set of subtour elimination constraints. We have tested our approach with $M = \{x : x_{ij} + x_{ji} \leq 1, 1 \leq i, j \leq n\}$ to avoid 2-cycles. An optimal solution (objective value 2,624) for this problem is shown in Fig. 3. If we delete the constraints of type 9 then the resulting problem A^* is comparable to an assignment problem where only $n - 1$ nodes are assigned. In [2] the first author analyzed the bipartite weighted matching problem with respect to slightly changed problems of the original problem. In one type of problem two nodes are deleted in the bipartite graph (one at each partition). The solution is of course a complete matching (an assignment) with $n - 1$ edges and therefore also a solution for A^* which can be computed in $O(n^3)$.

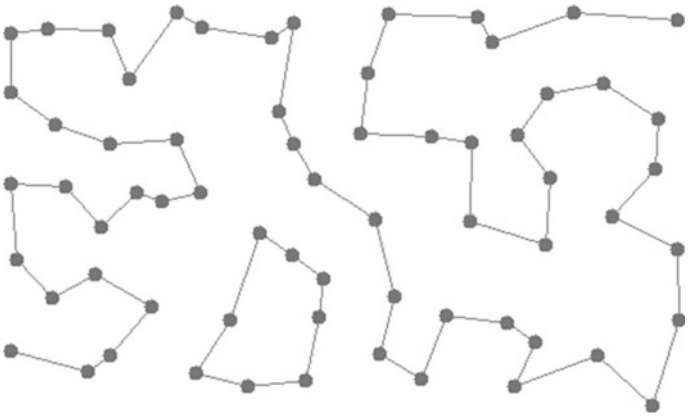


Fig. 3 Relaxation with $n - 1$ edges

Let $f(A')$ be the objective value of the above problem A'. OPT denotes the optimal solution of the original TSP and P' is any primal feasible solution. Then, of course we have

$$f(A') \leq OPT \leq f(P') \tag{11}$$

Theorem 1 *$f(P') - f(A')$ is an upper bound for the longest edge in an optimal solution of the TSP.*

Proof For any primal feasible solution P with objective value $f(P) \leq f(P')$ we claim:

If (i, j) is the longest edge in P then $c_{ij} \leq f(P') - f(A')$.

Suppose $f(P') - c_{ij} < f(A')$ then $P \setminus \{(i, j)\}$ is a feasible solution for problem A' with objective value $f(P) - c_{ij}$. Hence, $f(P) - c_{ij} \leq f(P') - c_{ij} < f(A')$ by our assumption. However, $f(A')$ was optimal and therefore we have a contradiction. This means that all edges longer than $f(P') - f(A')$ can not be in a better solution than P', in particular all these edges can not be in an optimal solution. □

In our example our best primal solution was 2,726 and the objective value of A' is 2,624. Therefore the difference 102 of these values is an upper bound for the longest edge in an optimal solution. This improves the value 2,352 computed via A^*. In Fig. 4 the edge (a, b) has length 104 and therefore this edge can not be in an optimal solution. All in all 3,542 edges (or 83 %) are longer than the computed bound and can be deleted.

Remark 1 There are TSP instances where the longest edge of the problem is in an optimal solution.

If all cities are on a semicircular then the longest edge (the diameter of the circular) is of course in the optimal solution. In this case our bound is useless.

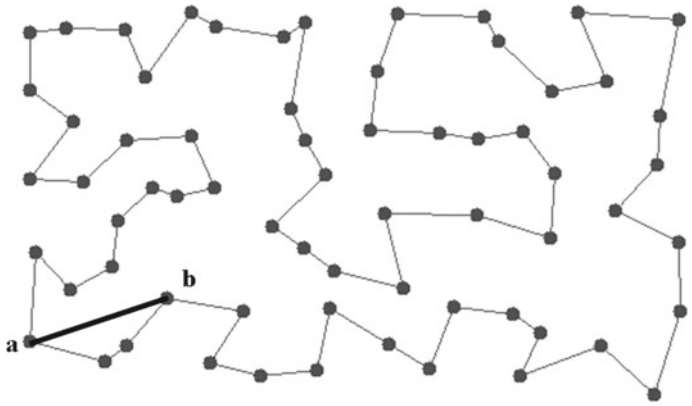


Fig. 4 Example for an edge to be deleted

Table 1 Upper bounds for longest edge

Instance	Cities	Length	Edges	Percentage
bay29	29	104	304	37
eil51	51	26	1,562	61
gr120	120	94	6,314	44
a280	280	130	32,788	42
att532	532	4,656	60,950	22

3 Computational Results

We have analyzed our approach with several instances in TSPLIB [5] where our set M was chosen to be $M = \{x : x_{ij} + x_{ji} \leq 1, 1 \leq i, j \leq n\}$. The first two columns denote the name and size of the problem.

The entries of the column "length" are the computed upper bounds for the respective instances. In the last two columns the number of edges longer than this bound and their percentage is given. This means for example for the drilling problem in instance a280 that 42 % of all edges are too long to be in an optimal solution. In all instances, computing the euclidian distances from the problem data takes more time than the computation of the LP-solution of A'. All primal feasible solutions were produced by the Lin-Kernighan heuristic [4].

Our computed bounds may be helpful computationally as they lead to potentially much sparser graphs to be considered in various algorithms.

References

1. Achatz, H., Kleinschmidt, P., & Paparrizos, K. (1991). A dual forest algorithm for the assignment problem. In P. Gritzmann, B. Sturmfels & V. Klee (Eds.), *Applied geometry and discrete mathematics. The Victor Klee festschrift* (pp. 1–10). Providence, R.I.: AMS (DIMACS'4).
2. Achatz, H. (1999). Sensitivity analysis of the bipartite weighted matching problem. In P. Kall (Ed.), *Operations Research Proceedings 1998. Selected Papers of the International Conference on Operations Research*, Zürich (pp. 135–141). Aug31–Sept 3, 1998. Berlin: Springer.
3. Applegate, D. L., Bixby, R. E., Chvátal, V., & Cook, W. J. (2007). *The traveling salesman problem. A computational story*. Princeton, NJ: Princeton University Press (Princeton series in applied mathematics).
4. Lin, S., & Kernighan, B. W. (1973). An effective heuristic algorithm for the traveling-salesman problem. *Operations Research*, *21*, 498–516.
5. Reinelt, G. (1991). TSPLIB—a traveling salesman problem library. *ORSA Journal on Computing*, *3*, 376–384.

Electricity Storage Systems and Their Allocation in the German Power System

Sonja Babrowski, Patrick Jochem and Wolf Fichtner

Abstract The need for daily electricity storage systems increases with the growing share of volatile renewable energy in the generation mix. Since the location of decentralized electricity generation (based on reneweable energy resource potentials) and electricity demand (depending on industrial facilities and population density) in Germany are geographically apart from each other, at the same time more electricity has to be transported. At certain times, this might challenge the transmission grid. Storage systems can be used for storing the surplus production of renewable energy and also help to prevent congestions in the grid. However, besides the technical feasibility there are economic criteria decisive for the installation of storage systems. These depend firstly on potential alternative technologies as gas turbines or the load shift potential of electric vehicles and secondly on the price development of storage systems. In order to estimate the future demand and the strategic allocation of daily storage systems in this context, expansion options for storage systems are implemented in the optimizing energy system model PERSEUS-NET-TS. This is a myopic material and energy flow model with an integrated nodal pricing approach. A mixed-integer optimization calculates the expansion and use of power plants in Germany until 2040 considering the DC restrictions of the transmission grid. Hence, the commissioning and allocation of storage systems in the German transmission grid is determined when the government target of 60 % renewable feed-in by 2040 is met. For this paper about every forth car in Germany is considered to drive electrically by 2040. When they are charged uncontrolled, directly after arrival the results are that by 2040 about 19 GW of storage systems are commissioned. Most are built closely to generation centers, but some are allocated close to bottlenecks in the transmission grid instead. When load shifting of the demand for electric mobility is allowed in terms of a controlled charging the required daily storage capacity could be reduced by more than half, so that only 8 GW are needed in 2040.

S. Babrowski (✉) · P. Jochem · W. Fichtner
Karlsruhe Institute of Technology (KIT), Institute for Industrial Production, Hertzstr. 16, 76187 Karlsruhe, Germany
e-mail: Sonja.Babrowski@kit.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_2,

1 Introduction

According to the objectives of the Federal Government in Germany 60 % of the gross electricity generation should be from renewable sources by 2040 [5]. Because of the volatile supply of wind and solar power, this is not going to work without adjustments of the transmission grid and either an additional thermal (reserve) power plant fleet or electricity storage systems. So far, electricity has mostly been generated at the time and place where it was needed. With the construction of large wind farms in the North and Baltic Sea this is going to change. In the future, generated electricity has to be harmonized with the demand in terms of place and time by transporting and storing it. Storage systems seem to be an alternative to the curtailment of renewables and could also serve as peak-load generation unit. Simultaneously, storage systems can be used for congestion management and thus lead to an improved utilization of the existing transmission capacities.

In the following, the commissioning and use of storage systems in the context of the future energy system in Germany is calculated with the energy system model PERSEUS-NET-TS. Possible alternative technologies such as gas turbines or load shifting through electric vehicles (EVs) are taken into account as well as restrictions due to the transmission grid.

2 The Energy System Model PERSEUS-NET-TS

PERSEUS-NET-TS is a bottom-up model of the German energy system. It is written in GAMS and uses the CPLEX solver. Depending on the setting it is solved by linear or mixed-integer programming. It is a myopic follow-up model of PERSEUS-NET [2] with a focus on daily storage systems. Besides the generation system a DC power flow model of the high and extra high voltage grid (220 and 380 kV) is integrated. Until 2040 PERSEUS-NET-TS calculates the commissioning and dispatch of the generation system at least for every fifth year. The model endogenously decides on commissioning coal, lignite, combined-cycle and gas power stations, as well as storage systems. These power plant extension options are modeled at most of the 442 network nodes depicted in the model and their configuration is based on [3]. About 550 modeled transmission lines connect these nodes. Larger power plants (over 100 MW) of the current generation system are directly assigned to specific grid nodes. The capacities of smaller power plants are aggregated by NUT3 regions (county level). Their cumulated capacities are assigned to the two closest grid nodes, inverse to the distance from the center of the region to them. In PERSEUS-NET-TS existing power plants are decommissioned 40 years after their commissioning. The conventional electricity demand is also calculated for each grid node based on the GDP and the number of inhabitants of adjacent NUT3 regions [2]. Besides this demand there is a demand for electric mobility integrated for each country based on forecasts made by the German Aerospace Center [6]. Accordingly 22 % of the

personal vehicles in Germany are going to be electric by 2030. This share remains stable until 2040 for the PERSEUS-NET-TS calculations. Depending on the settings this extra demand can either be charged controlled or uncontrolled directly after arriving at a charging opportunity at home or at the work place. When charged controlled the only restriction is that each day the needed amount has to be charged, but the model decides endogenously at which hour during the plug-in time. The driving force of the optimization is the exogenously given hourly electricity demand for a winter and a summer week. This hourly demand must be met at each grid node considering restrictions of the transmission grid and techno-economic constraints of the generation system. The electricity can be either generated at the grid node with the assigned power plants or transmitted over the grid from one of the neighboring nodes. The decision relevant expenditures are minimized for each of the considered periods. These expenditures include fuel costs and CO_2 certificate prices that are based on the World Energy Outlook 2012 [7]. Furthermore, the variable costs of the generation are considered, as well as costs for load changing of the thermal power plants (coal, lignite and gas-combined cycle plants). The fixed costs of the resulting power plant fleet in the current period are also added as well as investments for new power plants.

The commissioning of storage systems is allowed from 2020 on. The capital expenditures for storage systems are assessed to 1,000 EUR/kW in 2020 and are gradually reduced to 700 EUR/kW in 2040. Additionally, for battery storage systems there is a fixed ratio of installed capacity to storage volume (kW to kWh) of 1 to 5 assumed. This ratio is chosen according to the characteristics of daily storage systems in [1]. With the use of PERSEUS-NET-TS it is possible to determine the technology and the allocation of new generation plants. In order to prevent storage systems from storing and generating at the same time binary variables are needed for each hour and each storage. Thus, the model has to be solved via a mixed-integer optimization. With 336 considered hours (two weeks) for each period, the calculation time increases significant with every implemented storage option. Through the storing and generating at the same time, the efficiency of the storage systems can be used to "waste" electricity. This may make sense within the optimization to avoid the shutdown of thermal power plants with load changing costs or to meet the exogenously given targets for the renewable feed-in. To avoid this simultaneous bidirectional use of storage systems, we chose a two-step approach for this analyses (see Fig. 1). First, ideal storages with an efficiency of 100 % are implemented. In this case PERSEUS-NET-TS can be solved linearly with a relatively short computation time of about 9.5 h.[1] Each of the calculated 7 periods consists of about 2.7 million equations and 2.3 million variables. In this step, about 350 nodes distributed across Germany are provided with extension options for battery storage systems. In addition, a total of 30 pump storage power plants are considered, of which 10 are integrated as expansion option because they are currently in the planning phase.

[1] With the use of 6 threats on Windows Server 2008 R2 Enterprise, Intel(R) Xeon(R) CPU E5-1650 @ 3.20 GHz 3.20 GHz; 96 GB RAM; 64 Bit.

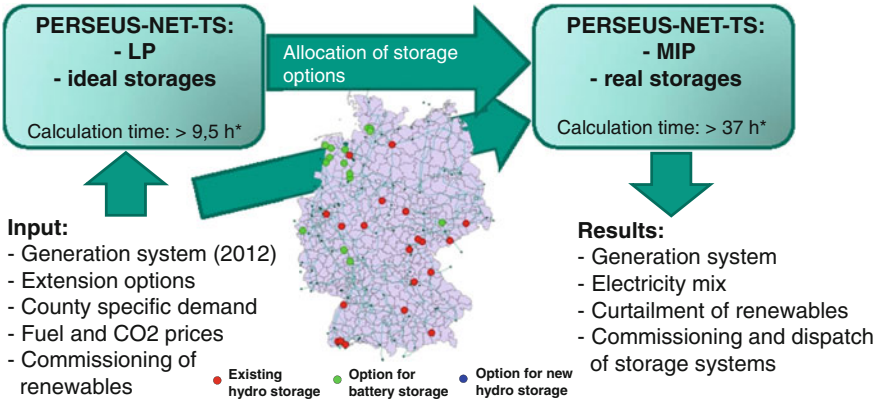


Fig. 1 Approach to reduce the needed binary variables

The resulting generation system of 2040 is analyzed in terms of the allocation of endogenously build storage systems. Only the grid nodes where ideal storage systems have been commissioned with the linear optimization are possible allocations for real storage systems in the following mixed-integer optimization. Storage systems in the mixed-integer optimization are modeled with an efficiency of about 80 %. The results presented below are based on this recalculation of PERSEUS-NET-TS as a mixed-integer problem with a calculation time of over 37 h.

3 Results

The development of the thermal generation system is a key result of PERSEUS-NET-TS. In 2040 a given demand of about 530 TWh has to be covered, of which 42 TWh occur due to electric mobility (about 8 %). The development of the renewables is exogenously given and has been derived from the German pilot study 2011 of the Federal Ministry of the Environment [4]. Overall, the installed capacity will increase in Germany until 2040 to about 224 GW, but at the same time the thermal power plant fleet is declining to about 47 GW (see Fig. 2, left). The commissioning of storage systems starts 2030 when 50 % renewables are targeted and peaks 2040 with an target of 60 % and battery prices of 700 EUR/kW, respectively 140 EUR/kWh. With about 19 GW storage systems represent about 8 % of the installed capacity by 2040. 12 GW consist of battery storage systems. Furthermore, considering only the endogenous installed power plants, i.e. the commissioning of new thermal power plants and storage systems, the commissioned storage systems represent about 40 % of the endogenously installed capacity until 2040 (see Fig. 2, right). According to the PERSEUS-NET-TS results most of the battery storage systems are built in the north-west near to the coast and thus close to the feed-in from offshore wind farms

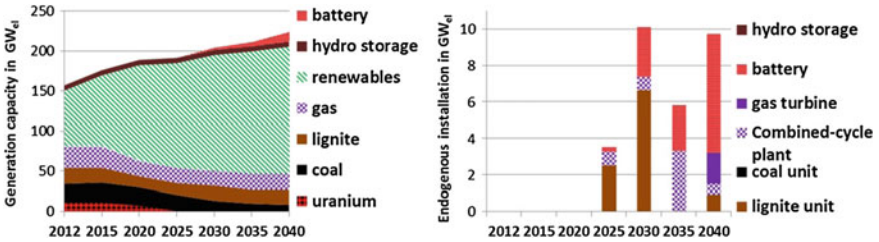


Fig. 2 Installed capacity (*left*) and endogenous commissioned capacity (*right*)

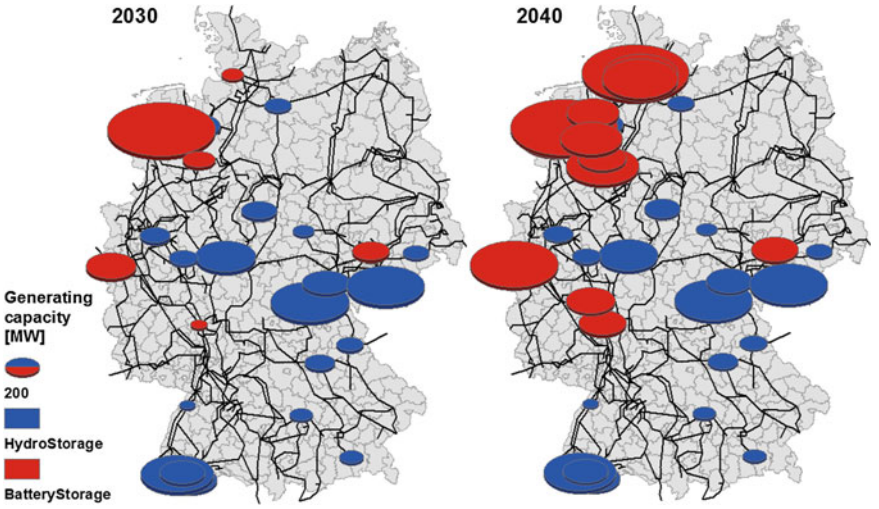


Fig. 3 Allocation of endogenously commissioned storage systems

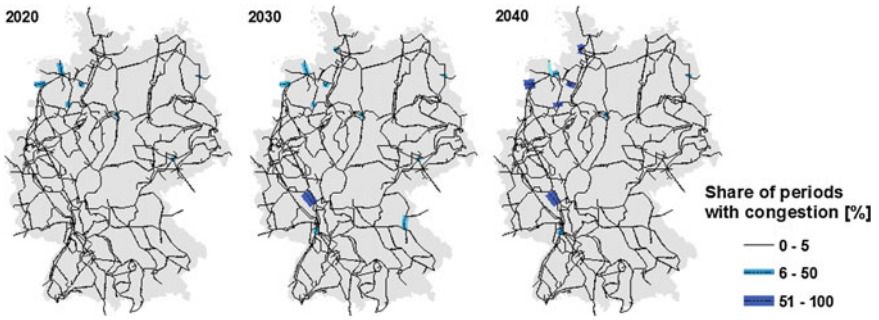


Fig. 4 Congestions in the transmission grid

in the North Sea (see Fig. 3). However, looking at the bottleneck in the transmission grid in the west of Germany (see Fig. 4) and the storage systems allocated on both sides of that bottleneck in 2040 (see Fig. 3) it becomes clear that storage systems are also used for grid management within the optimization.

In comparison to the results of an uncontrolled charging the need for storage systems decreases by 55 % to only 8 GW when the charging process of the EVs is controlled. In that case only about 2 GW battery storage systems are endogenously commissioned in the northwest.

4 Conclusion and Outlook

Through the implementation of storage options on transmission grid level in the energy system model PERSEUS-NET-TS we showed that the commissioning of storage systems in Germany is going to make sense considering an increasing renewable feed-in. According to the results the generation system should consist of about 8 % storage systems by 2040 (60 % renewables). These storage systems should be in general allocated close to the generation centers. Furthermore, the results indicate that a strategic allocation of storage systems might also help to prevent bottlenecks in the transmission grid. A second calculation showed that a big part of the needed flexibility could also be granted through a controlled charging of EVs instead. This alternative is certainly more economic assuming that the EV penetration rate is high enough and that the technology and the acceptance for an automatic controlled charging exists. However, interpreting these results it has to be kept in mind, that in favor to storage systems no net exchange with neighboring countries was allowed in the model and that on the other hand no stochastics were considered for the renewable feed-in. A more detailed analysis of the interaction between storage systems, renewable feed-in, power plants and the transmission grid is therefore going to be subject for further scientific work with PERSEUS-NET-TS.

References

1. Association for Electrical, Electronic and Information Technologies (VDE). (2012). *Energiespeicher für die Energiewende Speicherbedarf und Auswirkungen auf das Übertragungsnetz für Szenarien bis 2050*. Frankfurt.
2. Esser-Frey, A. (2012). Analysing the regional long-term development of the German power system using a nodal pricing approach. Dissertation, Karlsruher Institute of Technology (KIT), Karlsruhe.
3. German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU). (2010). Long-term scenarios and strategies for the deployment of renewable energies in Germany in view of European and global developments—Pilot study. Berlin.
4. German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU). (2011). Long-term scenarios and strategies for the deployment of renewable energies in Germany in view of European andglobal developments—Pilot study. Berlin.

5. German Federal Ministry of Economics and Technology (BMWi) and German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU). (2011). The Federal Government's energy concept of 2010 and the transformation of the energy system of 2011. Berlin.
6. Heinrichs, H. (2013). Analyse der langfristigen Auswirkungen von Elektromobilität auf das deutsche Energiesystem im europäischen Energieverbund unter Berücksichtigung von Netzrestriktionen. Dissertation, Karlsruher Institute of Technology (KIT), Karlsruhe.
7. International Energy Agency, (IEA). (2012). World energy, outlook.

Misspecified Dependency Modelling: What Does It Mean for Risk Measurement?

Theo Berger

Abstract Forecasting portfolio risk requires both, estimation of marginal return distributions for individual assets and dependence structure of returns as well. Due to the fact, that the marginal return distribution represents the main impact factor on portfolio volatility, the impact of dependency modeling which is required for instance in the field of Credit Pricing, Portfolio Sensitivity Analysis or Correlation Trading is rarely investigated that far. In this paper, we explicitly focus on the impact of decoupled dependency modeling in the context of risk measurement. We do so, by setting up an extensive simulation analysis which enables us to analyze competing copula approaches (Clayton, Frank, Gauss, Gumbel and t copula) under the assumption that the "true" marginal distribution is known. By simulating return series with different realistic dependency schemes accounting for time varying dependency as well as tail dependence, we show that the choice of copula becomes crucial for VaR, especially in volatile dependency schemes. Albeit the Gauss copula approach does neither account for time variance nor for tail dependence, it represents a solid tool throughout all investigated dependency schemes.

1 Introduction

Interdependencies between individual assets need to be captured to measure diversification effects and to precisely measure a single asset risk contribution on an aggregated portfolio level. Albeit, as Fantazzini [3] points out, the impact of misspecified marginals offsets the bias in dependency modeling on a portfolio level, precise dependency measurement represents a crucial information. For instance, a risk manager needs to know the effect of a hedged risk position on the overall portfolio risk. As well, correlation trading, the modeling of derivatives and measuring risk

T. Berger (✉)
University of Bremen, Wilhelm-Herbst-Str. 5, 28359 Bremen, Germany
e-mail: theoberger@uni-bremen.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_3,

diversification heavily depends on the information which are exclusively captured by dependency measurement.

So far there are only a few analysis which explicitly address the impact of dependency modeling on the measurement of Portfolio risk. Ane and Kharoubi [1] analyse the choice copula in the context of VaR forecasts, and show that inadequate dependency modeling explains up to 18 % of a VaR misspecification.[1] However, given the dominant impact stemming from the marginal return distributions, the separated impact of dependency modeling on aggregated portfolio risk in the absence of misspecified margins has not been explicitly investigated so far.

Thus, we add to the literature and set up an extensive simulation analysis accounting for realistic dependency scenarios such as time varying dependency and tail dependency. Both phenomena are discussed in a realistic as well as disproportionated environment. Further we investigate the dependency bias of modern dependency approaches on portfolio risk, in the absence of any bias caused by the modeling of marginal return distributions. More concrete, we generate samples with predefined margins, characterized by different dependency schemes and apply competing dependency models to forecast portfolio risk. By doing so, we are able to explicitly compare the forecasting bias caused by the applied dependency approaches in an applied risk measurement environment via out of sample analysis. Specifically, the simulation exercise should answer the question whether the choice of copula does affect the VaR performance when the data generating process is described by time varying conditional correlations or tail dependence.

The remainder is structured in the following way: Sect. 2 gives a brief overview about the relevant dependency approaches and Sect. 3 describes the setup of the simulation analysis. Section 4 gives the empirical results and Sect. 5 summarizes the results of this paper.

2 Methodology

2.1 Copulas

The copula approach is based on Sklar's Theorem [7]:
Let $X_1, \ldots, X_n$ be random variables, $F_1, \ldots, F_n$ the corresponding marginal distributions and H the joint distribution, then there exists a copula C: $[0, 1]^n \rightarrow [0, 1]$ such that:

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{1}$$

Conversely if C is a copula and $F_1, \ldots, F_n$ are distribution functions, then H (as defined above) is a joint distribution with margins $F_1, \ldots, F_n$.

[1] The analysis is based on applied loss functions in an empirical setup.

The Gaussian and t copula belong to the family of elliptical copulas and are derived from the multivariate normal and t distribution respectively.

The setup of the Gaussian copula is given by:

$$C^{Ga}(x_1, \ldots, x_n) = \Phi_\rho(\Phi^{-1}(x_1), \ldots, \Phi^{-1}(x_n)), \tag{2}$$

$$= \int_{-\infty}^{\Phi^{-1}(x_1)} \cdots \int_{-\infty}^{\Phi^{-1}(x_n)} \frac{1}{2(\pi)^{\frac{n}{2}} |\rho|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} z^T \rho^{-1} z\right) dz_1 \ldots dz_n \tag{3}$$

whereas Φ_ρ stands for the multivariate normal distribution with correlation matrix ρ and Φ^{-1} symbolizes the inverse of univariate normal distribution.

Along the lines of the Gaussian copula, the t copula is given by:

$$C^t(x_1, \ldots, x_n) = t_{\rho,v}(t_v^{-1}(x_1), \ldots, t_v^{-1}(x_n)), \tag{4}$$

$$= \int_{-\infty}^{t^{-1}(x_1)} \cdots \int_{-\infty}^{t^{-1}(x_n)} \frac{\Gamma\left(\frac{v+n}{2}\right)}{\Gamma\left(\frac{v}{2}\right) (v\pi)^{\frac{n}{2}} |\rho|^{\frac{1}{2}}} \left(1 + \frac{1}{v} z^T \rho^{-1} z\right)^{-\frac{v+n}{2}} dz_1 \ldots dz_n, \tag{5}$$

in this setup $t_{\rho,v}$ stands for the multivariate t distribution with correlation matrix ρ and v degrees of freedom (d.o.f.). t_v^{-1} stands for the inverse of the univariate t distribution and v influences tail dependency. For $v \to \infty$ the t distribution approximates a Gaussian.

In contradiction to the elliptical copulas, the Clayton copula belongs to the group of Archimedean copulas and is given by:

$$C^{Clayton}(x_1, x_2) = \left(\max\{x_1^\theta + x_2^\theta - 1, 0\}\right)^{\frac{1}{\theta}}, \tag{6}$$

with $\theta \in [-1, \infty) \backslash \{0\}$. Note that the Clayton copula describes stronger dependence in the negative tail than in the positive, for $\theta \to \infty$ the Clayton copula describes comonoticity, and for $\theta \to 0$ independence.

Another popular Archimedean copula is represented by the Gumbel copula which, in contradiction to the Clayton copula, exhibits higher dependence in the positive tail than in the negative. The copula is given by:

$$C^{Gumbel}(x_1, x_2) = \exp\left(-[(-ln\ x_1)^\delta + (-ln\ x_2)^\delta]^{\frac{1}{\delta}}\right), \tag{7}$$

with $\delta \in [1, \infty)$. Analogue to the Gumbel copula, we get comonoticity for $\theta \to \infty$ and independence for $\theta \to 0$.

As well we introduce the Frank copula as defined by Nelson (1999) which is given by:

$$C^{Frank}(x_1, x_2) = -\frac{1}{\theta} ln\left(1 + \frac{(e^{-\theta x_1} - 1)(e^{-\theta x_2} - 1)}{e^{-\theta} - 1}\right), \tag{8}$$

for $\theta \in \mathbb{R}\backslash\{0\}$.

Due to the fact that estimating parameters for higher order copulas might be computationally cumbersome, all parameters are estimated in a two step maximum likelihood method given by Joe [5]. This approach is also known as inference for the margins (IFM). The two steps divide the log likelihood into one term incorporating all parameters concerning univariate margins and into one term involving the parameters of the chosen copula. Thus, this method enables us to explicitly isolate the dependency modeling from fitting the univariate marginals.

2.2 VaR

In order to make the results of the competing copula approaches comparable, we translate the figures into a VaR universe, so that we are able to evaluate the properties of different copulas within a realistic risk measurement framework. Generally, VaR is defined as the quantile at level α of the distribution of portfolio returns:

$$VaR_\alpha = F^{-1}(\alpha) = \int_{-\infty}^{VaR_\alpha} f(r)dr = P(r \leq VaR_\alpha). \tag{9}$$

So that, the respective quantiles are direct functions of the variances, which enables us to directly translate the quantiles of the estimated portfolio variances into VaR figures. Let α be the quantile, H_t the covariance matrix and w the vector of portfolio weights, then VaR at time t is given by: $VaR_t = -\alpha\sqrt{w' H_t w}$ for both normal and t distributions. For instance the 99 % VaR of PF return y_t represents the empirical 1 % quantile of the variance.

3 Simulation Design

The aim of the simulation exercise is to analyze the impact of dependency modeling apart from the choice of optimal marginal distribution. Further, we explicitly address the isolated impact stemming from dependency modeling on quantile forecasts from two angles:

Three different dependency scenarios (weak/medium/strong) are investigated for each copula (Table 1).

Once the sample covering 1001 observations is generated, we use the introduced copula approaches to forecast Value-at-Risk. At this, the forecast is based on 1000

Table 1 Archimedean copulas: simulated scenarios

Copula	Low dependency	Medium dependency	High dependency
Clayton	$\theta = 0,5$	$\theta = 1,5$	$\theta = 2,5$
Gumbel	$\theta = 1$	$\theta = 3$	$\theta = 5$
Frank	$\theta = 10$	$\theta = 20$	$\theta = 30$

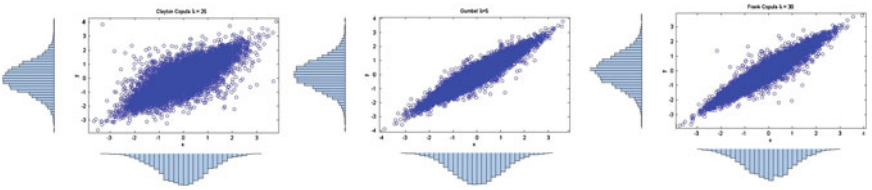


Fig. 1 1001 normally distributed return observations generated by Clayton copula ($\theta = 25$), Gumbel copula ($\theta = 5$) and Frank Copula ($\theta = 30$)

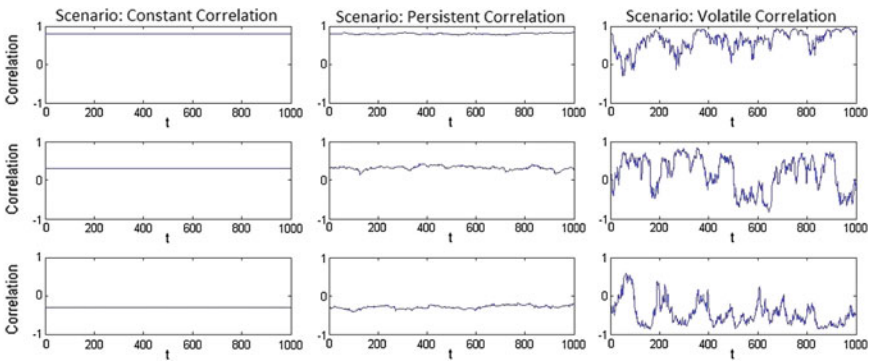


Fig. 2 Correlation scenarios

observations and can be evaluated by the 1001th one. This approach enables us to investigate the VaR performance and thus the compatibility of the competing copula approaches when the samples are not in line with the underlying assumptions. Figure 1 illustrates examples of the investigated scenarios. 10.000 scenarios for each dependency modification are simulated and evaluated. Secondly, to take account for financial time series specific properties, we generate normally distributed return series characterized by time varying conditional correlations (see Fig. 2) to investigate the consequences of time varying dependency schemes on the applied copula approaches. Again, for each scenario 10.000 return series covering 1001 observations are generated and the competing copula models are evaluated by the VaR forecasting performance regarding the 1001th observation.

4 Simulation Results

To sum it up, the empirical VaR backtesting performance for all investigated scenarios is given in Table 2. Obviously, given that dependency is modeled via time varying linear correlation coefficient, the elliptical copulas outperform the Archimedean copulas in terms of empirical VaR performance. The Clayton copula underestimates risk when the data generating process is not a Clayton copula whereas mixed evidence can be reported for both Gumbel and Frank copula. However, both approaches, are not able to adequately capture tail dependence generated via Clayton copula. By comparing both Gauss and t copula, the overall performance of the t copula is slightly more precise when it comes to forecast VaR.[2] According to the return series generated by Archimedean copulas, mixed evidence can be reported. For 95 % VaR forecasts, the Gumbel and Frank copula applied to returns generated via Clayton copula result in an inappropriately high number of misspecifications[3] and thus both dependency approaches would be rejected by standard statistical VaR backtesting.[4] Interestingly, the Gumbel and Frank copula lack in capturing tail dependence generated by Clayton copula and vice versa. However, given that the empirical backtesting performance of all the other investigated models are "statistically" acceptable, leads us to conclude that misspecified dependency modeling has an impact on the rejection of a VaR model. Thus, given that the rejection rate is impacted by the choice of the applied copula approach, our findings are twofold:

- From a pregulatory perspective, the classical gauss and t copula seem to be an appropriate choice for all investigated dependency scenarios. According to our results, it is the Gauss and t copula which mainly result in the "second best" solution[5] when the returns are generated by different copulas. Thus, the higher parametrisation of the competing copula approaches does not lead to more precise dependency measurement and hence more accurate VaR failure rate. Further, due to the higher parameterisation, the Archimedean copulas lack in terms of preciseness when the underlying sample does not exhibit the characteristics of the applied copulas. Moreover, having in mind that Gauss copulas do neither account for time varying dependency structure nor for tail dependence, we show that the parsimonious approach leads to acceptable VaR figures throughout all investigated scenarios.
- However, from an institutional point of view, it is not only the rejection rate which is relevant but also the absolute size of VaR. If we analyse both, the rejection rate as well as the absolute amount of VaR forecast, we favour the model which results in the lowest amount of VaR forecast, given that the empirical failure rate is in line with the expectations. For time varying linear dependency, again, the elliptical

[2] We applied the CPA test proposed by Giacomini and White (2006) to prove this fact. Results are available upon request.

[3] The empirical backtesting performance would get rejected by statistical backtesting criteria, "conditional coverage", by Christoffersen (1998).

[4] Results are available upon request.

[5] The "first best" solution is always the original model.

Table 2 Empirical misspecification performance, 95 % and VaR forecasts

Scenario	G Cop (%)	t Cop (%)	Clayton (%)	Gumbel (%)	Frank (%)
95 % VaR					
Elliptical	4,97	4,99	4,03	4,68	5,04
Clayton	6,10	6,09	5,16	7,05	6,37
Gumbel	4,71	4,70	4,60	4,82	4,77
Frank	5,07	5,07	4,89	5,44	4,92

copulas do outperform the Archimedean approaches, since they result in the lowest VaR values and show an acceptable empirical failure rate. Along the lines of the linear dependency scenarios, the elliptical copulas also represent the (second-) best choice for VaR forecasts for samples which are generated by Archimedean copulas. The Archimedean copulas heavily depend on the assumptions of the underlying samples, so that Frank copulas adequately capture tail dependency generated by Gumbel copula (and vice versa) whereas both Frank and Gumbel fail to capture characterstics generated by Clayton copula.

5 Conclusion

Albeit the main impact on multivariate portfolio VaR stems from the choice of marginal return distributions, the adequate modeling of dependency needs to be considered in order to achieve an appropriate VaR performance. So that, when it comes to the impact of dependency modeling on VaR forecasts, the choice of copula is crucial.

Based on the given extensive simulation analysis covering different dependency scenarios and triggered by the comparison of competing copula approaches, we conclude that the investigated elliptical copulas do outperform the Archimedean copulas due to more precise VaR forecasts. Thus, having in mind that both the Gauss and t copula are straightforward to apply to multivariate asset portfolios comprising three or more assets, we strongly recommend the application of elliptical copulas in the context of VaR forecasts.

References

1. Ane, T., & Kharoubi, C. (2003). Dependence structure and risk measure. *The Journal of Business*, *76*(3), 411–438.
2. Christoffersen P. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862.
3. Fantazzini, D. (2009). The effects of misspecified marginals and copulas on computing the value at risk: A Monte Carlo study. *Computational Statistics and Data Analysis*, *53*(6), 2168–2188.

4. Giacomini, R. and H. White (2006). Tests of Conditional Predictive Ability. *Econometrica*, *74*(6), 1545–1578.
5. Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. In L. Ruschendorf, B. Schweizer, M. D. Taylor (Eds.), *Distributions with fixed margins and related topics* (Vol. 28, pp. 120–141). IMS Lecture Notes Monograph Seriex.
6. Nelson, R.B. (1990). *An Introduction to Copulas*. New York: Springer.
7. Sklar, C. (1959). *Fonctions de repartition a n dimensions et leurs marges* (Vol. 8, pp. 229–231). Publicationss de Institut Statistique de Universite de Paris.

Primal MINLP Heuristics in a Nutshell

Timo Berthold

Abstract Primal heuristics are an important component of state-of-the-art codes for mixed integer nonlinear programming (MINLP). In this article we give a compact overview of primal heuristics for MINLP that have been suggested in the literature of recent years. We sketch the fundamental concepts of different classes of heuristics and discuss specific implementations. A brief computational experiment shows that primal heuristics play a key role in achieving feasibility and finding good primal bounds within a global MINLP solver.

1 Introduction

Optimization problems that feature, at the same time, nonlinear functions as constraints and integrality requirements for the variables are arguably among the most challenging problems in mathematical programming. This article gives an overview on existing heuristic approaches to find good feasible solutions for these so-called *MINLPs*.

Definition 1 (*MINLP*) A mixed integer nonlinear program (MINLP) is an optimization problem of the form

$$\begin{aligned} &\min c^{\mathsf{T}} x \\ &\text{s.t.} \quad g_i(x) \leq 0 \quad \text{for all } i \in \mathscr{M} \\ &x_j \in \mathbb{Z} \qquad\quad \text{for all } j \in \mathscr{I}, \end{aligned}$$

T. Berthold (✉)
Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany
e-mail: berthold@zib.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_4,

where
$\mathcal{I} \subseteq \mathcal{N} := \{1, \ldots, n\}$ is the index set of the integer variables, $c \in \mathbb{R}^n$, and $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i \in \mathcal{M} := \{1, \ldots, m\}$.

There are many subclasses of MINLPs; in this article, we will be particularly concerned with the following: *convex MINLPs*, for which all constraint functions $g_i, i \in \mathcal{M}$, are convex, *mixed integer quadratically constrained programs (MIQCPs)*, for which all constraint functions are quadratic, *mixed integer linear programs (MIPs)*, for which all constraint functions are linear, *nonlinear programs (NLPs)*, for which all variables are continuous, and *linear programs (LPs)*, for which the constraints are linear and all variables are continuous.

For MIPs, it is well-known that general-purpose *primal heuristics* like the feasibility pump [1, 19, 20] are able to find high-quality solutions for a wide range of problems. A *primal heuristic* is, roughly speaking, an incomplete algorithm that aims at finding high-quality feasible solutions quickly. In general, it is neither guaranteed to be successful, nor does it provide any additional information, such as a dual bound on the solution quality.

For MINLPs, research in the last five years has shown an increasing interest in primal heuristics [7–9, 11, 12, 14, 17, 22, 24, 25]. The goal of this article is to provide a brief overview on the cited work. We focus on methods that have been developed for the application inside a global solver such as BARON, BONMIN, COUENNE, or SCIP. In such an environment, it is often worth sacrificing success on a number of instances for a significant saving in average running time. One way to do so are "fast fail" strategies that take the most crucial decisions in the beginning and in a defensive fashion such that if the heuristic aborts, it will not have consumed much running time. Furthermore, we restrict ourselves to primal heuristics that have been specifically developed and tested for MINLPs; we do not cover the manifold ideas to apply metaheuristics to global optimization problems.

We partition our survey by the main concepts on which the reviewed algorithms are based. Nonlinear extensions of the feasibility pump [19] are discussed in Sect. 2, large neighborhood search heuristics are introduced in Sect. 3, other ideas, such as rounding and diving, are treated in Sect. 4. Section 5 presents a computational evaluation of primal heuristics implemented within the MINLP solver SCIP.

2 Feasibility Pumps

The fundamental idea of all Feasibility Pump [19] algorithms is to construct two sequences of points that hopefully converge to a feasible solution of a given mathematical programming problem. One sequence consists of points that are feasible for a continuous relaxation (e.g., an NLP relaxation of an MINLP), but possibly integer infeasible. The other sequence consists of points that are integral (for the integer variables), but might violate the imposed constraints. The next point of one sequence is always generated by minimizing the distance to the last point of the

other sequence, using different distance measures in both cases (e.g., the ℓ_1 and the ℓ_2 norm). We refer to the process of constructing an integral point from a constraint feasible point as the *rounding step* and to the process of finding a new point that fulfills the continuous relaxation as the *projection step*.

Bonami et al. [11] and Bonami and Gonçalves [12] present the first two versions of a Feasibility Pump for MINLPs. Both teams of authors consider convex MINLPs and implement their ideas in Bonmin et al. [10].

The paper [12] is probably the closest to the original Feasibility Pump for MIPs. It performs a simple rounding to the nearest integer in the rounding step and solves a convex NLP relaxation with an ℓ_1 objective for the projection step.

In [11], the authors suggest using an ℓ_2 norm as objective for the projection step. The most significant difference to [12], however, is the implementation of the rounding step. Instead of performing an instant rounding to the nearest integer, they solve a MIP relaxation based on an outer approximation [16] of the underlying MINLP. This has an important effect w.r.t. the main weakness of Feasibility Pump algorithms: cycling. For convex MINLPs, it is always possible to avoid cycling by adding a no-good cut to the auxiliary MIP. The particular difficulty addressed by D'Ambrosio et al. in [14] is that of handling the nonconvex NLP relaxation when adapting the algorithm of [11] to nonconvex constraints. The authors suggest using a stochastic multi-start approach, feeding the NLP solver with multiple randomly generated starting points, and solving the NLP to local optimality. In the event that this does not lead to a feasible solution, a final NLP is solved, in which the integer variables are fixed and the original objective is re-installed on the continuous variables. To avoid cycling, their algorithm provides the MIP solver with a tabu list of previously used solutions.

3 Large Neighborhood Search

The main idea of *large neighborhood search* (LNS) is to define a neighborhood of "good" solution candidates centered at a particular reference point—typically the incumbent solution. The neighborhood is explored by solving an auxiliary MINLP, which is constructed by restricting the feasible region of the original MINLP by additional constraints and variable fixings. LNS is a common paradigm for MIP heuristics, e.g., RINS [15], which defines a neighborhood by fixing variables which coincide in the incumbent and the LP optimum, or Local Branching [18], which searches the neighborhood of solutions that differ in at most k variables from the incumbent.

Bonami and Gonçalves describe an extension of the RINS heuristic to convex MINLPs [12]. They use an optimum of the NLP relaxation as a second reference solution besides the incumbent.

Nannicini, Belotti, and Liberti introduce a Local Branching heuristic for nonconvex MINLPs [24]. It solves a MIP that is derived from a linear relaxation of the original MINLP, the integrality constraints, and a Local Branching constraint.

Subsequently, an NLP local search is performed by fixing the integer variables to the values from the Local Branching MIP's incumbent (which is not necessarily feasible for the original MINLP) and solving the resulting continuous problem.

In [9], Berthold et al. suggest a generic way of generalizing LNS heuristics from MIP to MINLP, for the first time presenting nonlinear versions of Crossover [4, 26] and the DINS [21] heuristic.

Berthold presents RENS [7], an LNS algorithm that optimizes over the set of feasible roundings of a relaxation solution. To this end, integer variables that take an integral value in the relaxation solution are fixed to that value, for others, the bounds are changed to the two nearest integers.

In [8], Berthold and Gleixner introduce *Undercover*, an LNS start heuristic for MINLP that explores a linear subproblem which is obtained by fixing as small a subset of variables as possible. The set of variables to be fixed is determined by solving a vertex covering problem. Although general in nature, this approach works best for MIQCPs.

The RECIPE algorithm described in [22] falls into the category of variable neighborhood search heuristics: it iteratively explores different neighborhoods, updating the neighborhood definition after each iteration.

4 Rounding, Diving, and MIP Heuristics

Rounding, diving, and propagation heuristics are kind of "folklore": Most solvers and many custom codes use them, but there are few publications on this topic.

Bonami and Gonçalves present computational results for NLP-based diving heuristics [12]. Their algorithm solves a convex NLP relaxation, fixes several variables (with variable selection rules referred to as Fractional Diving and Vectorlength Diving in [4]), and iterates this process. They further tested solving a final sub-MINLP as soon as all fractional variables exclusively belong to linear constraints. Mahajan et al. [23] suggest a diving algorithm that uses quadratic programming relaxations.

Nannicini and Belotti present *iterative rounding* [25], which is a mixture of diving and variable neighborhood search. It solves a series of auxiliary MIPs to generate integer points near an initial optimal solution of an NLP relaxation. In each iteration, the feasible region of the MIP gets contracted further by outer approximation and no-good cuts.

A popular approach for solving MINLPs is to use an outer approximation generated by linearization of convex constraints and linear underestimation of nonconvex constraints. Having an outer approximation at hand, one might employ MIP primal heuristics to the outer approximation LP plus the integrality constraints. In particular for heuristics that are computationally very cheap, such as rounding and propagation heuristics [3], this is a valid strategy. Applying MIP heuristics to such a "MIP relaxation" typically produces points that are integral, valid for the LP outer approximation, but violate one or more nonlinear constraints. Such points are natural candidates for an NLP local search as it is, e.g., described in [17, 22, 24]: the integer variables

are fixed to their value in the (infeasible) reference solution and the resulting NLP is solved to local optimality.

5 Computational Results

To evaluate the impact of primal heuristics on the performance of a global MINLP solver, we conducted a computational experiment in which we compare the performance of the MINLP solver SCIP [2] when running with and without primal heuristics. We used SCIP version 3.0.1 compiled with SOPLEX 1.7.1 [28] as LP solver and IPOPT 3.11 [27] as NLP solver. SCIP does not run all of the described algorithms by default. It features Undercover, nonlinear versions of RENS and Crossover, an NLP local search, and many MIP heuristics, including a Feasibility Pump (for an overview, see [5]). As a test set, we chose the MINLPLIB [13], excluding instances which feature nonlinear functions that SCIP 3.0.1 cannot handle, e.g., trigonometric functions. The results were obtained on a cluster of 64 bit Intel Xeon X5672 CPUs at 3.20 GHz with 12 MB cache and 48 GB main memory, running an OPENSUSE 12.3 with a GCC 4.7.2 compiler. We imposed a time limit of 1 h.

Similar to the situation in MIP, the impact of primal heuristics on the overall running time was negligible. Both versions differed by less than one percent in shifted geometric mean. Furthermore, both variants solved 170 of the 252 test instances to optimality. The major difference occurs when considering the primal bound. For those instances which could not be solved within the time limit, the SCIP version without heuristics found a feasible solution in 35 cases, the one using primal heuristics in 58. The primal bound at termination was better for 48 instances when using primal heuristics, only for two instances it was worse. Consequently, the average *primal integral* [6] of both runs differed by about 50 %.

References

1. Achterberg, T., & Berthold, T. (2007). Improving the feasibility pump. *Discrete Optimization, Special Issue*, *4*(1), 77–86.
2. Achterberg, T. (2009). SCIP: Solving constraint integer programs. *Mathematical Programming Computation*, *1*(1), 1–41.
3. Achterberg, T., Berthold, T., & Hendel, G. (2012). Rounding and propagation heuristics for mixed integer programming. In D. Klatte, H.-J. Luthi, & K. Schmedders (Eds.), *Operations research proceedings 2011* (pp. 71–76). Berlin Heidelberg: Springer.
4. Berthold, T. (2006). Primal heuristics for mixed integer programs. *Diploma thesis*, Technische Universität Berlin.
5. Berthold, T. (2008). Heuristics of the branch-cut-and-price-framework SCIP. In J. Kalcsics & S. Nickel (Eds.), *Operations research proceedings 2007* (pp. 31–36). New York: Springer.
6. Berthold, T. (2013). Measuring the impact of the primal heuristics. *Operations Research Letters*, *41*(6), 611–614.
7. Berthold, T. (2014). RENS: the optimal rounding. *Mathematical Programming Computation*, *6*(1), 33–54.

8. Berthold, T., & Gleixner, A. M. (2014). Undercover: A primal MINLP heuristic exploring a largest sub-MIP. *Mathematical Programming*, *144*(1–2), 315–346.
9. Berthold, T., Heinz, S., Pfetsch, M. E., & Vigerske, S. (2011). Large neighborhood search beyond MIP. In L. D. Gaspero, A. Schaerf, & T. Stutzle (Eds.), *Proceedings of the 9th Metaheuristics International Conference (MIC 2011)* (pp. 51–60).
10. Bonami, P., Biegler, L., Conn, A., Cornuéjols, G., Grossmann, I., Laird, C., et al. (2008). An algorithmic framework for convex mixed integer nonlinear programs. *Discrete Optimization*, *5*, 186–204.
11. Bonami, P., Cornuéjols, G., Lodi, A., & Margot, F. (2009). A feasibility pump for mixed integer nonlinear programs. *Mathematical Programming*, *119*(2), 331–352.
12. Bonami, P., & Gonçalves, J. (2012). Heuristics for convex mixed integer nonlinear programs. *Computational Optimization and Applications*, *51*, 729–747.
13. Bussieck, M., Drud, A., & Meeraus, A. (2003). MINLPLib a collection of test models for mixed-integer nonlinear programming. *INFORMS Journal on Computing*, *15*(1), 114–119.
14. D'Ambrosio, C., Frangioni, A., Liberti, L., & Lodi, A. (2012). A storm of feasibility pumps for nonconvex MINLP. *Mathematical Programming*, *136*, 375–402.
15. Danna, E., Rothberg, E., & Pape, C. L. (2004). Exploring relaxation induced neighborhoods to improve MIP solutions. *Mathematical Programming*, *102*(1), 71–90.
16. Duran, M. A., & Grossmann, I. E. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, *36*(3), 307–339.
17. Extending a CIP framework to solve MIQCPs. In J. Lee, & S. Leyffer (Eds.), Mixed integer nonlinear programming. *Volume 154 of The IMA volumes in mathematics and its applications* (pp. 427–444). Springer.
18. Fischetti, M., & Lodi, A. (2003). Local branching. *Mathematical Programming*, *98*(1–3), 23–47.
19. Fischetti, M., Glover, F., & Lodi, A. (2005). The feasibility pump. *Mathematical Programming*, *104*(1), 91–104.
20. Fischetti, M., & Salvagnin, D. (2009). Feasibility pump 2.0. *Mathematical Programming Computation*, *1*, 201–222.
21. Ghosh, S. (2007). DINS, a MIP improvement heuristic. In M. Fischetti & D. P. Williamson (Eds.), *Proceedings of 12th International IPCO Conference on Integer Programming and Combinatorial Optimization* (Vol. 4513 of LNCS, pp. 310–323). Springer.
22. Liberti, L., Mladenovic, N., & Nannicini, G. (2011). A recipe for finding good solutions to MINLPs. *Mathematical Programming Computation*, *3*, 349–390.
23. Mahajan, A., Leyffer, S., & Kirches, C. (2012). Solving mixed-integer nonlinear programs by QP-diving. Preprint ANL/MCS-2071-0312, Argonne National Laboratory, Mathematics and Computer Science Division.
24. Nannicini, G., Belotti, P., & Liberti, L. (2008). A local branching heuristic for MINLPs. ArXiv e-prints.
25. Nannicini, G., & Belotti, P. (2012). Rounding-based heuristics for nonconvex MINLPs. *Mathematical Programming Computation*, *4*(1), 1–31.
26. Rothberg, E. (2007). An evolutionary algorithm for polishing mixed integer programming solutions. *INFORMS Journal on Computing*, *19*(4), 534–541.
27. Wächter, A., & Biegler, L. (2006). On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, *106*(1), 25–57.
28. Wunderling, R. (1996). Paralleler und objektorientierter Simplex-Algorithmus. PhD thesis, Technische Universität Berlin.

Layout Optimisation of Decentralised Energy Systems Under Uncertainty

Valentin Bertsch, Hannes Schwarz and Wolf Fichtner

Abstract We present a modelling approach to support the layout optimisation of decentralised energy systems composed of photovoltaic (PV) panels and heat pumps with thermal storage capabilities. The approach integrates the simulation-based generation of model input on the basis of publicly available meteorological data and the subsequent optimisation. Selected results concerning the choice of an appropriate storage size are presented for an illustrative decentralised energy system.

1 Introduction

In the context of the ongoing transformation of the electricity generation system with an emphasis on renewables and low-carbon generation as well as the implementation of smart grid technologies, intelligent home energy management approaches making use of load flexibilities are discussed increasingly often. Especially, photovoltaic (PV) systems in combination with heat pumps and thermal storages have attracted attention in the recent past. The dimensioning of the individual components, such as the storage size, has an immediate impact on the system's economic performance. When modelling such systems using linear programming (LP) techniques, a variety of input data, subject to different sources of uncertainties, needs to be provided. Thus, we present an approach integrating modules for (a) simulating input data, such as solar irradiation or temperature profiles, by a stochastic process, (b) transforming these initial profiles to consistent sets of PV generation and heat demand profiles and (c) using the generated profiles in an optimisation.

This paper is structured as follows: The problem and its LP formulation are described in Sect. 2. In Sect. 3, we present our modelling approach focussing on

V. Bertsch (✉) · H. Schwarz · W. Fichtner
Institute for Industrial Production (IIP), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: valentin.bertsch@kit.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_5,

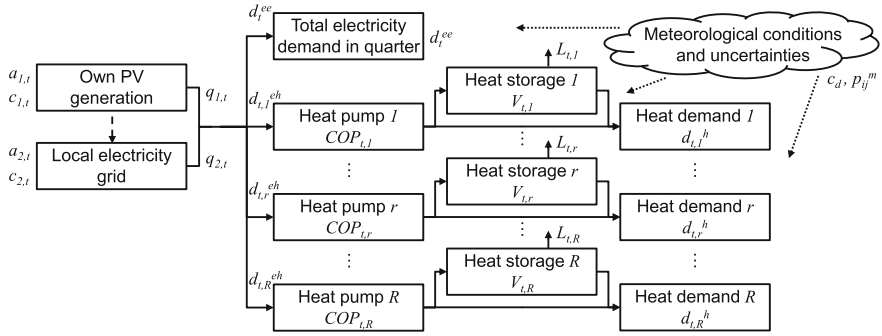


Fig. 1 Illustrative setup of the decentralised energy system

the stochastic simulation modules. Illustrative results are shown in Sect. 4. Finally, Sect. 5 summarises the main findings and indicates needs for further research.

2 Problem Description

Within decentralised energy systems, we focus on a residential quarter including several groups of multi-family or row houses. The setup is illustrated in Fig. 1. The energy is provided by own PV panels or the local grid. Fossil fuels are not used. Energy for room heating and hot water, represented by one heat demand profile per group of houses, is provided by heat pumps and storages for each group. The electricity demand beyond heating and hot water is considered as an aggregated profile for the whole quarter. If the PV generation exceeds the demand within the quarter, it can be fed into the grid. The quantities in Fig. 1 are explained in Table 1. As for (large) energy systems models, the system illustrated in Fig. 1 can be modelled as a classical LP problem. We define the target function as the minimisation of the total system expenses within each year, see (1). The storage size is not modelled as a continuous variable but varied exogenously on a discrete basis according to storages available on the market. Equations (2–5) represent the most important constraints. Equation (2) ensures that the used quantity $q_{f,t}$ of each electricity source does not exceed its availability $a_{f,t}$ at any time t. For the PV generation, $a_{f,t}$ is the fluctuating generation profile. Equation (3) guarantees that demand and supply are balanced at all times t. Subsequently, Eq. (4) represents the storage possibility of heat, i.e. the main flexibility in the system. Constraint (5) ensures that the storage volume can only vary within the given boundaries.

$$\min \sum_{t} \sum_{f} q_{f,t} \cdot c_{f,t} \tag{1}$$

Table 1 Nomenclature

Parameters		Variables		Indices	
d_t^{ee}	Electricity demand for electrical usage at time t	$V_{t,r}$	Volume of heat storage of building group r at time t	t	Time index indicating the hour $t \in \{1, \ldots, 8{,}760\}$ of the year
$d_{t,r}^{h}$	Heat demand of building group r at time t	$d_{t,r}^{eh}$	Used electricity for heating of building group r at time t	f	'Fuel' type index ($f = 1$: el. from PV; $f = 2$: el. from grid)
$a_{f,t}$	Availability of 'fuel' f at time t	$q_{f,t}$	Used quantity of 'fuel' f at time t	r	Index of the R building groups within the quarter: $r \in \{1, \ldots, R\}$
$c_{f,t}$	Costs of 'fuel' f at time t			m	Index $m \in 1, \ldots, 12$ indicating the month
$COP_{t,r}$	COP value of the heat pump of building group r at time t			d	Index indicating the day
$V_r^{\max}/V_r^{\min}$	max/min volume of heat storage of building group r				
$L_{t,r}$	Losses of heat storage of building group r at time t				
R	Amount of building groups within quarter				
Stochastic modelling quantities					
c_d	Random variable for the cloudiness on day d	p_{ij}^m	Transition probability: $p_{ij}^m = p(c_d = j \mid c_{d-1} = i)$		

subject to

$$0 \leq q_{f,t} \leq a_{f,t} \quad \forall t \, \forall f \tag{2}$$

$$\sum_{f} q_{f,t} = d_t^{ee} + \sum_{r} d_{t,r}^{eh} \quad \forall t \tag{3}$$

$$COP_{t,r} \cdot d_{t,r}^{eh} + V_{t-1,r} = d_{t,r}^{h} + V_{t,r} + L_{t,r} \quad \forall t \, \forall r \tag{4}$$

$$V_r^{\min} \leq V_{t,r} \leq V_r^{\max} \quad \forall t \, \forall r \tag{5}$$

3 The Developed Approach to Layout Optimisation

In order to solve the optimisation problem described by (1–5), manifold input data is needed. E.g., assumptions on the development of electricity prices ($c_{2,t}$) and profiles for the PV generation ($a_{1,t}$), the electricity (d_t^{ee}) and the heat ($d_{t,r}^{h}$) demand need to be available. However, this input data is subject to many different uncertainties which need to be addressed adequately in the layout planning process. To generate the necessary input data and to account for the associated uncertainties, we propose an integrated approach supporting the generation of consistent ensembles of load and solar PV profiles, i.e. it includes the fundamental relationships between weather and load as well as PV generation. These profiles are used in the subsequent optimisation. Hence, our approach includes three subsystems (see Fig. 2):

- The Weather Simulation Subsystem (WSS)
- The Demand and Supply Subsystem (DSS)
- The Economic Evaluation Subsystem (EES)

3.1 The Weather Simulation Subsystem

The main task of the WSS consists in the generation of solar irradiation and temperature profiles considering their stochastic nature. For the modelling of solar irradiation variations, several authors proposed Markov processes. Focussing on the long-term variations, Amato et al. [1] model daily solar irradiation using a Markov model. Focussing on the short-term variations in a high time resolution, Morf [4] proposes a Markov process aimed at simulating the dynamic behaviour of solar irradiation. However, the input data is often not available in the required granularity. Since our focus is on layout planning, our approach needs to take into account both, the short-term as well as the long-term variations, since both of these may affect the choice of the storage size. Hence, we suggest a two-step approach.

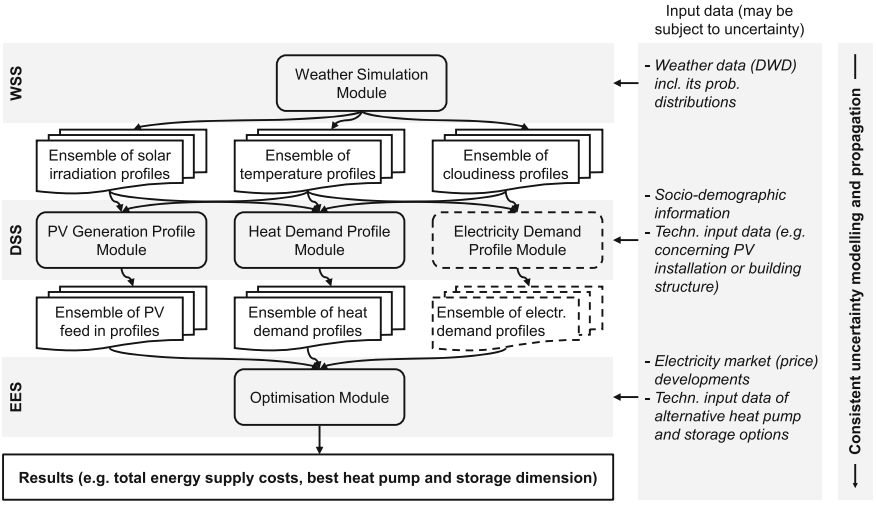


Fig. 2 Conceptual structure of the integrated modelling approach

In the *first step*, we model daily values in order to account for the long-term variations. Having an immediate impact on the heat demand as well as the efficieny of the solar PV panels, we need temperature profiles in addition to the solar irradiation profiles. We therefore go one step back and simulate daily meteorological conditions, particularly the cloudiness, on whose basis the values for daily solar irradiation and average daily temperature are derived under consideration of seasonality information (e.g. the months). In this sense, our approach goes beyond Ehnberg and Bollen [2] who simulate solar irradiation on the basis of cloud observations using a discrete Markov process. However, neither do they use the cloud observations for simulating consistently compatible temperature profiles nor do they introduce a monthly component in their markov process. Besides these differences, our approach is similar to the one proposed by [2]. The daily cloudiness $c_d \in \{0, \ldots, 8\}$ is considered in oktas in our Markov process, describing how many eighths of the sky are covered by clouds [3]. The transition probabilities p_{ij}^m for each month are derived on the basis of publicly available weather data from Germany's National Meteorological Service (DWD), which are available for a variety of locations across Germany for periods of often more than 50 years. Overall, a backtesting of the monthly Markov process shows good results, not only concerning the bandwidth and distribution of the average yearly cloudiness but also concerning the standard deviation of the daily cloudiness values.

In a *second step*, a stochastic process is used to generate hourly profiles on the basis of the daily simulation results of step 1. Besides the results of step 1, the process is based on hourly solar irradiation and temperature data, which was collected for Karlsruhe, Germany, over a period of 4 years. While 4 years would be a short period

for understanding long-term variations, this period provides a valuable basis for modelling short-term fluctuations of solar irradiation and temperature.

3.2 The Demand and Supply Subsystem

The DSS's task is the transformation of the meteorological profiles into PV supply and heat demand profiles to be used in the subsequent optimisation. Theoretically, the DSS could also be used to generate electricity demand profiles but since we only consider the electricity demand on the quarter level, i.e. the total demand of approx. 70 households, we use the so-called 'standard load or H0 profile' as our analysis shows a strong convergence of the aggregate household load towards the H0 profile even for numbers of households much lower than 70. For the solar generation, a physical PV model has been developed on the basis of [5]. Concerning the heat load, a reference load profile approach is currently implemented in the DSS. The approach is based on the VDI4655 guideline [6] and uses the temperature profiles as an important input. In the long run, we envisage to replace the reference load approach by a physical model in order to achieve a higher accuracy.

3.3 The Economic Evaluation Subsystem

The ESS uses the profiles processed by the DSS as input and allows for carrying out economic optimisations on the basis of Eqs. (1–5). In the current state, PLEXOS for Power Systems®, a tool for power market modelling and simulation,[1] is used as optimisation module within the ESS. In the long run, we envisage to develop an own optimisation module tailored to the specific needs of our problem.

4 Illustrative Results

Figure 3 illustrates an example of the results, which can be produced by our approach. We applied our approach for meteorological data of Karlsruhe, Germany. The underlying demand and supply profiles correspond to an average meteorological year, the long-term variations are not yet considered. Moreover, a flat electricity tariff of 25 ct/kWh is assumed. It should be noted that already today, more attractive tariffs are available for heat pumps, especially, bearing in mind that not the individual households but the whole quarter of approx. 70 households can negotiate tariffs with the electricity supplier. The absolute IRR values should therefore not be overrated.

[1] See: http://www.energyexemplar.com. We thank Energy Exemplar for the provision of the software and their support.

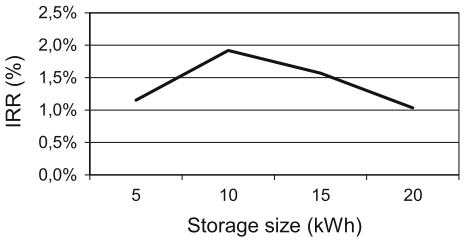


Fig. 3 Internal rate of return (IRR) for different storage sizes (Assumptions: 25 years technical lifetime; approx. 240 EUR additional invest per kWh additional storage capacity)

For a flat electricity tariff, however, a relative comparison shows a maximal IRR for a storage size of 10 kWh in our example.

5 Conclusions and Outlook

An approach to support the layout planning of decentralised energy systems has been presented. The approach has been applied to a residential quarter including approx. 70 households and allows for choosing an appropriate storage size. The approach now provides the basis for further analyses of the economic profitability of such systems as well as service related business models.

Possible enhancements of the approach include the implementation of a physical model to generate heat demand profiles, the development of a tailor-made optimisation module and the increase of the time resolution. Moreover, the uncertainties associated with the input data of the different modules and their impact on the results needs to be analysed and visualised in more detail.

References

1. Amato, U., Andretta, A., Bartoli, B., Coluzzi, B., Cuomo, V., Fontana, F., et al. (1986). Markov processes and Fourier analysis as a tool to describe and simulate daily solar irradiance. *Solar Energy*, *37*(3), 179–194.
2. Ehnberg, J. S., & Bollen, M. H. (2005). Simulation of global solar radiation based on cloud observations. *Solar Energy*, *78*(2), 157–162.
3. Jones, P. (1992). Cloud-cover distributions and correlations. *Journal of Applied Meteorology*, *31*, 732–741.
4. Morf, H. (1998). The stochastic two-state solar irradiance model (stsim). *Solar Energy*, *62*(2), 101–112.
5. Ritzenhoff, P. (2006). Erstellung eines Modells zur Simulation der Solarstrahlung auf beliebig orientierte Flachen und deren Trennung in Diffus- und Direktanteil. Berichte des Forschungszentrum Jülich, 2600, Überarbeitete Fassung.
6. VDI4655. (2006). Reference load profiles of single-family and multi-family houses for the use of CHP systems. Verein Deutscher Ingenieure (VDI), Guideline 4655.

Analysis of Micro–Macro Transformations of Railway Networks

Marco Blanco and Thomas Schlechte

Abstract A common technique in the solution of large or complex optimization problems is the use of *micro–macro* transformations. In this paper, we carry out a theoretical analysis of such transformations for the track allocation problem in railway networks. We prove that the *cumulative rounding* technique of Schlechte et al. satisfies two of three natural optimality criteria and that this performance cannot be improved. We also show that under extreme circumstances, this technique can perform inconveniently by underestimating the global optimal value.

1 Introduction

It is often the case in discrete optimization problems coming from applications that the data is too complex to be tractable by an efficient algorithm. However, much of the information in this precise (also called *microscopic*) model is not necessary to obtain a very good feasible solution. A common technique is to derive a simplified *macroscopic* model by aggregating the structures of the microscopic model, find a good solution to the macroscopic model, and retranslate it to the original problem. This idea has been used in diverse settings. In [1], an algorithm for solving linear programs exactly solves a sequence of increasingly detailed LPs until the desired degree of precision is reached. In [2], an algorithm for solving a dynamic progam over a large state space is described. A sequence of coarse DPs is solved, and the complexity/level of detail increases gradually. Reference [3] surveys aggregation and disaggregation techniques for optimization problems. This research was mostly influenced by Schlechte et al. [5], where a micro–macro transformation is used for

M. Blanco (✉) · T. Schlechte
Zuse Institute Berlin, Takustr. 7, 14195 Dahlem, Berlin, Germany
e-mail: blanco@zib.de

T. Schlechte
e-mail: schlechte@zib.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_6,

solving the track allocation problem for railway networks (see [4, 5] for a precise definition), which is the problem considered in this paper. One of the main difficulties in developing an efficient micro–macro algorithm for this problem is choosing a reasonable time discretization. That is, given a time unit δ in the microscopic model, we seek to find a larger unit Δ for the macroscopic model and then determine the input times of the macroscopic model in multiples of Δ. It is on this last step that we will focus next. Given a microscopic running time of some route on a macroscopic track, the most natural choice is to round it to a close multiple of Δ. Rounding down can lead to infeasibilities, while rounding up all running times leads to an unnecessary increase in the optimal value. Therefore, a combination of both seems to be the best strategy. In this context, we consider the *cumulative rounding* method introduced by Schlechte et al. in [5]. This method consists of rounding up the running times along each route in order of traversal, until the total "lost" time accumulated is at least the time corresponding to the track currently considered, at which point we round down this running time and iterate.

While it is possible to give upper bounds on the overestimation error of the total time needed to traverse each route, the impact of this rounding on the originating network optimization problem as a whole has not been studied. The paper is structured as follows. In Sect. 2 we describe the general problem, the motivation and the goals of micro–macro transformations. In Sect. 3 we define three optimality criteria for a rounding strategy for the track allocation problem. We prove that the cumulative rounding strategy is optimal with respect to two of these criteria and that no strategy satisfies all three of them. Finally, in Sect. 4 we show an instance in which cumulative rounding yields a macroscopic value that is smaller than the microscopic optimum and whose solution is impossible to translate back to the original model without losing a significant factor. This shows the difficulty of acheiving global optimality or near-optimality.

2 Our Setting

We consider a general minimization[1] problem P_δ based on a time discretization δ with $k\delta = \Delta, k \in \mathbb{Z}, k > 0$. The problem P_Δ results from rounding all times of P_δ to multiples of Δ with respect to alternate rounding strategies. Let us consider the trivial rounding down ($\lfloor\ \rfloor$) and up ($\lceil\ \rceil$). Then for the optimal values v, we have:

$$v(P_\Delta^{\lfloor\rfloor}) \leq v(P_\delta) \leq v(P_\Delta^{\lceil\rceil})$$

On the one hand the solution of $P_\Delta^{\lceil\rceil}$ can be re-transformed, i.e., we maintain the orders of the trains and retranslate the departure and arrivals w.r.t. δ, to a feasible solution of P_δ retaining the same objective value or obtaining a better one. On

[1] In case of the track allocation problem we want to schedule a fixed number of trains on a network within a minimum time horizon.

the other hand $v(P_{\Delta}^{\lfloor\rfloor})$ only provides in general a valid lower bound. Thus, we can guarantee some solution quality provided by the lower bound $v(P_{\Delta}^{\lfloor\rfloor})$.

3 Optimality Criteria

While the ultimate objective in the track allocation problem is to find a microscopic solution of optimal or near-optimal value, it is in general not clear how to obtain a feasible microscopic solution from a macroscopic solution such that the objective value does not increase. For that reason, we will try to judge the quality of a transformation by comparing the values of the obtained macroscopic and microscopic solutions. There are several (often conflicting) possibilities of defining an "optimal" rounding algorithm, and it is not obvious which of them should be considered. Here we consider three very natural optimality criteria:

1. **Global optimality**: The total time is not underestimated and the corresponding (overestimating) error is minimal.
2. **Route-wise optimality**: The total time on each individual route is not underestimated and the corresponding (overestimating) error is minimal.
3. **Local optimality**: The overestimating error on any subroute $(j_m, j_{m+1}, \ldots, j_{m+n})$ of a route r is less than Δ.

The no-underestimating condition guarantees that we can obtain feasible solutions. The first two conditions are self-explaining and the third condition guarantees that the approximation is good on a local level, i.e., on intervals.

In this section we prove that the cumulative rounding technique satisfies the last two properties.

Theorem 1 *For the track allocation problem, a rounding strategy is route-wise optimal if and only if on every route j it rounds up the traversal times corresponding to exactly* $\left\lceil \frac{\sum_{j\in D} \widehat{t}_j^{\,r}}{\Delta} \right\rceil$ *tracks.*

Proof In the same setting as above, let r be a route. For every track j in the route, let t_j^r be the time (in units of δ) needed to traverse j, and let $\widehat{t}_j^{\,r} \equiv t_j^r \pmod{\Delta}$. If for this track we decide to round up, the (overestimating) error will be $\Delta - \widehat{t}_j^{\,r}$, while if we round down, the (underestimating) error is $\widehat{t}_j^{\,r}$. Let J be the set of tracks in route r, let $U \subset J$ (the set of tracks for which we round up) and $D = J \backslash U$ (the tracks for which we round down). Now, the total overestimating error is

$$\varepsilon^r = \sum_{j\in U} (\Delta - \widehat{t}_j^{\,r}) - \sum_{j\in D} \widehat{t}_j^{\,r} = |U|\Delta - \sum_{j\in J} \widehat{t}_j^{\,r}.$$

Since $\sum_{j\in J} \widehat{t}_j^{\,r}$ is independent of the choice of U and D, the total error depends only on the cardinality of U. By the non-overestimating property, we are looking for a

set U of minimal cardinality such that ε^r is nonnegative and minimal. Clearly, this is achieved by choosing U with $|U| = \left\lceil \frac{\sum_{j \in D} \widehat{t}_j^r}{\Delta} \right\rceil$. □

Corollary 1 *The cumulative rounding strategy is route-wise optimal and locally optimal.*

Proof The authors of [5] have proven that on each route, the total error caused by cumulative rounding is in the interval $[0, \Delta)$. By the proof of the previous theorem, this is a minimizer and thus the strategy is route-wise optimal.

To prove local optimality, let us consider a subroute $r^1 = (j_m, \ldots, j_{m+n})$. We can picture this subroute as the difference between subroutes $r^2 = (j_1, j_2, \ldots, j_{m+n})$ and $r^3 = (j_1, j_2, \ldots, j_{m-1})$. As before, let us denote by ε^r the overestimating error of a subroute r. By the result in [5] we just mentioned above, we have $0 < \varepsilon^{r^1} < \Delta$ and $0 < \varepsilon^{r^2} < \Delta$. Suppose $\varepsilon^{r^3} > \Delta$. Then, we clearly have $\varepsilon^{r^2} = \varepsilon^{r^1} + \varepsilon^{r^3} > \Delta$, which is a contradiction. □

Theorem 2 *There exists no rounding strategy that satisfies all three described optimality criteria.*

Proof Let us consider the following network, with $\Delta = k\delta$ for some $k \geq 3$:

On this network, let us consider trains 1 and 2 traveling from A to D, and train 3 traveling from D to A. We are interested in minimizing the time until the last train arrives at its destination. We assume that for every track, the headway time corresponding to two trains in the same direction is Δ. Similarly, the headway time corresponding to two trains in opposite directions along track j is $t_j + \Delta$. Suppose trains can not stop at intermediate stations and there are no restrictions on the departure or arrival times. A feasible and in fact optimal solution is to let trains 1, 2 and 3 leave their initial stations at times 0, Δ and δ, respectively. As trains 1 and 2 go from B to C in one direction, train 3 goes from C to B in the opposite direction without violating the headway constraints. The time until the last train (train 2) arrives is $5\Delta + \delta$. Suppose we have a route-wise and locally optimal strategy. Let us consider r^1, the route corresponding to train 1. By route-wise optimality, we know that exactly one traversal time is rounded down. If this time corresponds to either track AB or track CD, we know that the remaining two tracks form a subroute with an overestimating error of Δ, which contradicts local optimality. Without loss of generality, the same reasoning applies to routes r^2 and r^3, so the resulting macroscopic network is given by the numbers below the arcs on Fig. 1.

Let t_j denote the microscopic time for each train on track j and T_j the corresponding macroscopic time. By choice of the microscopic headway times, the macroscopic headway times are still Δ and $T_j + \Delta$. Since now the tracks between B and C are of time Δ, the previous solution is no longer feasible. In fact, now the optimal solution is to let trains 1 and 2 go from A to D, and let train 3 depart only after the other two have arrived at D. This gives a total time of 12Δ, which is more than double the time needed in the microscopic instance.

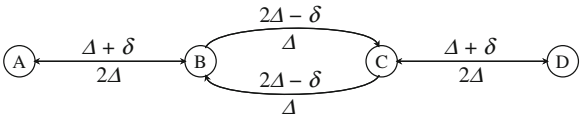


Fig. 1 The numbers *above* and *below* the *arcs* represent, respectively, the microscopic and macroscopic times for the corresponding tracks, assuming all trains have uniform speed

Table 1 Results for transformations of the optimization problem described in the proof of Theorem 2

Description	Discretization	Rounding technique	Optimal value
Original problem	δ	–	$5\Delta + \delta$
Approximation	Δ	$\langle\,\rangle$	12Δ
Feasible solution	δ	$\langle\,\rangle$	$10\Delta + 2\delta$
Feasible solution	Δ	$\lceil\,\rceil$	7Δ
Feasible solution	δ	$\lceil\,\rceil$	$5\Delta + \delta$
Lower bound	Δ	$\lfloor\,\rfloor$	4Δ

We use $\langle\,\rangle$ to denote any strategy that is route-wise- and locally optimal

Applying the conservative approach (rounding up all running times), we would get a total time of 7Δ as optimum, which is more than the microscopic optimal value but significantly smaller than 12Δ. Since the conservative rounding gives a smaller total time we can conclude that the considered strategy does not satisfy global optimality. □

While the previous proof shows that the conservative rounding strategy gives a better macroscopic total time, it is not immediately clear what the corresponding microscopic times are. If we take the solution given by cumulative rounding or a similar strategy and translate it back to the microscopic model, we obtain a total time of $10\Delta + 2\delta$, which is exactly double of the optimal time. We summarize these results in Table 1. Let us also note that we can easily make the macroscopic instance infeasible while keeping the original feasible. For example, we could require for all trains to arrive at their destinations at time 6Δ or before.

4 A Paradoxical Instance

In the previous section we saw some drawbacks to the cumulative rounding strategy, but we also proved that it is impossible to improve it to a globally optimal strategy while keeping both of its optimality properties. In this section, we will give an instance such that the macroscopic optimal value is much better than the microscopic optimal value. This shows that even if we relax the optimality requirement in the global optimality condition, the non-underestimating condition is not necessarily satisfied.

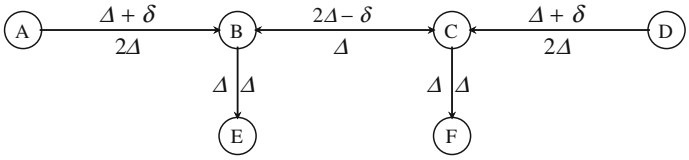


Fig. 2 The numbers *above* (*below*) and to the *left* (*right*) of the *arcs* represent the microscopic (macroscopic) times for the corresponding tracks

Furthermore, this hints that guaranteeing non-underestimation on a global level in general may be very hard. Consider the network with two trains in Fig. 2.

Here, train 1 has to go from A to F and train 2 from D to E. The only headway times of interest are those corresponding to track BC. They are defined as $t_{BC} + \Delta$. Trivially, an optimal solution is to let train 1 depart at time 0 and let train 2 depart when train 1 is about to reach C (to be precise, at time $3\Delta - \delta$). In this solution, train 2 arrives to its destination at time $7\Delta - \delta$.

As in the previous example, the macroscopic headway times of interest are now $T_{BC} + \Delta$. Letting train 1 depart at time 0 and train 2 at time 2Δ, the last train arrives at time 6Δ. Clearly, this objective value is impossible to attain in the microscopic problem.

References

1. Gleixner, A. M., Steffy D. E., & Wolter, K. (2012). Improving the accuracy of linear programming solvers with iterative refinement. In: *Proceedings of ISSAC 12* (pp. 187–194). New York: ACM.
2. Raphael, C. (2001). Coarse-to-fine dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(12), 1379–1390.
3. Rogers, D. F., Plante, R. D., Wong, R. T., & Evans, J. R. (1991). Aggregation and disaggregation techniques and methodology in optimization. *Operations Research*, *39*(4), 553–582.
4. Schlechte, T. (2012). Railway track allocation—models and algorithms. *PhD thesis*. Technische Universität Berlin, Berlin.
5. Schlechte, T., Borndörfer, R., Erol, B., Graffagnino, T., & Swarat, E. (2011). Micro–macro transformation of railway networks. *Journal of Rail Transport Planning and Management*, *1*(1), 38–48.

A Human Resource Model for Performance Optimization to Gain Competitive Advantage

Joachim Block and Stefan Pickl

Abstract Human resources are one of the most important assets of any organization. This is especially true in the ongoing information age where a transformation from blue to white collar work takes place. Implementing efficient and effective human resource management (HRM) policies can result in a sustained competitive advantage which will contribute to the long-term success of an organization. Even though empirical studies show some evidence that HRM policies influence organizational performance, the mechanisms at work are still not uncovered completely. We present a model of an organization's human resources from a holistic perspective. Using system dynamics and agent-based modeling and simulation, we integrated two usually distinct fields of HRM research: the micro and the macro perspective. First numerical simulation results indicate a promising approach to unlock the causal chain between HRM policies and organizational performance. Our research should not only contribute to fill an existing research gap but also help to identify HRM policies which enable management to optimize the performance of an organization.

1 Introduction

The modern world of today depends on well working organizations of all kinds. They are drivers for innovation and welfare. An organization is a goal directed social entity which is linked to the external environment and designed as deliberately structured and coordinated activity system [5]. In the transformation process from blue to white

J. Block (✉)
Department of Computer Science, COMTESSA,
University of the German Armed Forces, Munich, Germany
e-mail: archibald@ieee.org

S. Pickl
Department of Computer Science,
University of the German Armed Forces, Munich, Germany
e-mail: stefan.pickl@unibw.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_7,

collar work, human resources are not just a cost factor but a key strategic asset of any organization [7]. Recruiting and retaining better people and integrating their specific talents in better processes can result in a competitive advantage [3].

Therefore, human resource management (HRM) is one of the most important management tasks. By implementing HRM policies, management intends to increase organization's overall performance in order to guarantee its long-term success. However, there is no golden key how to organize the HRM system.

2 Bridging Macro and Micro HRM

The implementation of sustained and value creating HRM policies depends on a profound understanding of the mechanisms at work in the social system of the organization. A lack of such an understanding holds the probability that HRM activities will not contribute to performance improvement but causes the opposite, on the long-run.

One important challenge in HRM research is to unlock the causal chain between HRM activities and organizational performance [16]. The identification of this link holds some potential for maximizing the impact of HRM on organizational performance [20]. Research on this topic

> would benefit from a fuller integration of the micro (focused on designing recruitment, selection, performance appraisal, rewards systems, etc., as well as individual employee response to those systems) and macro (focused on the strategy formulation and implementation processes) domains [9, p. 422].

An integrated approach differs significantly from the traditional HRM research that is dominated by either a macro-level view or a micro-level view [20]. However, both views are two sides of the same coin.

Our research aims to fill the existing gap. We present a holistic and integrated human resource model of an organization which draws on the strengths of both perspectives. By simulating different HRM policies with the model, value creating and sustainable HRM policies can be identified. This should not only result in performance optimization of the whole organization but also in a sustained competitive advantage.

3 An Integrated Model of the Organization

Our model regards an organization as what it is: a multilevel system [10]. Main components of the model are two perspectives found in HRM research (see e.g. [20]):

1. The *workforce perspective* representing the macro or top-down view of the organization.
2. The *individual perspective* representing the micro or bottom-up view of the organization.

The *workforce perspective* is taken by top management or strategic management. Here, the flow of staff between different grades and the organization's environment is considered. Organizations as whole or bigger groups of staff are in focus. HRM policies are applied at this level to control the flow of staff. The wider field of workforce or manpower planning addresses this level (see e.g. [1]).

In contrast, the *individual perspective* considers individuals or smaller groups of individuals and their relations. This is the micro-level view of the organization. It delivers detailed data on individual or group behavior. HRM practices, usually applied by line managers, aim to influence individual behavior at this level.

Even though HRM policies are applied at the *workforce perspective* to increase organizational performance, it is the individuals which perform and not the organization [10]. Actions taken on the top level have to and do in fact impact behavior of the individual staff member [11]. On the other hand, decisions of employees can cause disruptions on the workforce level. An example is a promising staff member who leaves the organization because of better job opportunities elsewhere. His action causes a new flow to occur on the workforce level (e.g. entry of a new employee to fill the gap). Hence, *workforce perspective* and *individual perspective* not only are tightly connected but nested.

In the first steps of the modeling process the *workforce perspective* and the *individual perspective* are modeled as two subsystems. By using different modeling techniques, individual specifics of the macro and micro perspectives are adequately considered. The integration of both subsystems into a holistic system closes the modeling process.

3.1 The Workforce Subsystem

Stocks and flows are core elements of the *workforce subsystem*. Stocks represent different grades or positions in an organization, depicted for example in an organization chart [5]. They have a defined maximum carrying capacity. A certain number of employees is assigned to one of these stocks at a time. The number of grades or positions in the organization usually is much lower than the number of employees. Grades and positions are personnel categories. These categories differ from each other by certain attributes like responsibility, wage, requirements, etc.

Three kinds of flows can be identified in an organization [8]: the recruitment flow, the internal personnel flows between different personnel categories (among others promotion flows), and the wastage flow. New members enter the organization into certain grades or positions (recruitment). During their professional life, they move through different grades or positions (internal) until they leave the organization (wastage).

Due to the nature of being a top-level or strategic perspective, modeling is restricted to aggregated information. System dynamics (SD) is suited to study the dynamic behavior of the stock and flow workforce system (see e.g. [18] or [14]). It enables us to model feedbacks and nonlinearities that are inherent to the system [19].

3.2 The Individual Subsystem

In the *individual subsystem*, individual members of the organization are modeled as agents by the use of the correspondent modeling technique: agent-based modeling and simulation (ABMS). Agents are autonomous and self-directed individuals which live in and interact with an environment [12]. They are governed by certain rules and are able to change these rules. Agents have attributes and show behavior [15].

As we are interested in performance optimization, we build upon the AMO theory [4]. According to this theory, performance p of an individual i is a function of his or her ability (A), motivation (M), and opportunity (O):

$$p_i = f(A_i, M_i, O_i) \tag{1}$$

Though, ability, motivation, and opportunity are the core attributes in our ABMS model. We conducted a literature review in order to find parameters which influence these three elements and built a dependency graph from these information. Motivation is for example influenced by time spent in the actual position and by promotion prospects. On the other hand, ability depends among others on work experience and on training effort. Ability, motivation, and opportunity are subject to change as time goes by. The dependency graph reveals the existence of a feedback loop between ability and motivation. Such loops can be easily modeled by SD.

Besides modeling agents' attributes by SD we take use of this method for behavioral aspects, as well. Consistent to [6], relevant behavior is: performance, turnover, and absenteeism. In contrast to short-term absenteeism, turnover means that a member will leave the organization permanently. Attributes influence agents' behavior. For instance, a low level of motivation increases turnover probability.

3.3 The Integrated System

The ABMS model of the *individual perspective* and the SD model of the *workforce perspective* are connected. Integrating both perspectives means integrating SD and ABMS [17].

By flows being an indicator for promotion opportunities, the w*orkforce perspective* affects agents' behavior. On the other hand, behavior of individual agents influences the *workforce subsystem*. Turnover results in new flows to occur and the entrance of a new agent into the system to fill the gap. Furthermore, in case of absenteeism the overall workload has to be distributed among the other agents. As a result, agents' attributes change. To implement the depicted model we use AnyLogic® [2] which supports message passing between the two subsystems.

In addition, we implement some parameters that are external to the two subsystems but significantly impact behavior of the whole system. The external market conditions for example influence turnover rates. While the external market is out

of managerial control, some other parameters indeed are in. The level of training, gratification and job design can be adjusted by management intervention. Furthermore, the implementation of different promotion rules and adjustments of the grade or position structure on the workforce level are possible. These actions impact prospects for personal growth of all agents and thus behavior of the system as a whole.

Our model enables management to simulate different policies by changing model parameters. Thereby, long-term effects on overall performance can be evaluated and negative side effects identified. Underlying assumption is that overall performance P of the organization is a function of the performance of all agents:

$$P = g(p_1, p_2, \ldots, p_n) \tag{2}$$

with n being the number of individuals in the organization.

4 Discussion

We configured our model with data from a German public sector organization. Miscellaneous simulation runs show a system behavior which resembles past experiences in Germany's public administration [13]. By applying different HRM policies we studied the impact on organizational performance. First results show for example significant differences in turnover rates and performance between a policy which solely considers the oldest employee for promotion and a promotion policy which is based on performance.

Our hybrid SD and ABMS model seems to offer a promising way to unlock the causal chain between HRM policies and organizational performance. However, the results presented in this paper should be regarded as a starting point for further research in the combined micro–macro view within HRM rather than the end of a journey. There is still a long road to go. Especially, more effort has to be invested in model validation and verification. The theoretical framework will be extended and further aspects will be taken into consideration. One aspect will be about the implementation of not yet considered human attitudes. Last but not least, the calculation of organization's overall performance needs some critical reflections, as well.

References

1. Bartholomew, D. J., Forbes, A. F., & McClean, S. I. (1991). *Statistical Techniques for Manpower Planning* (2nd ed.). Chichester: Wiley.
2. Borshchev A. and Filippov A. (2004). From System Dynamics and Discrete Event to Practical Agent Based Modeling: Reasons, Techniques, Tools. Paper presented at the 22nd International Conference of the System Dynamics Society, Oxford, UK.
3. Boxall, P. (1996). The strategic HRM debate and the resource-based view of the firm. *Human Resource Management Journal, 6*(3), 59–75.

4. Boxall, P., & Purcell, J. (2011). *Strategy and Human Resource Management* (3rd ed.). Hampshire: Palgrave Macmillan.
5. Daft, R. L. (2009). *Organization Theory and Design* (10th ed.). Mason: South Western CENGAGE Learning.
6. Gardner, T. M., Moynihan, L. M., Park, H. J., & Wright, P. M. (2001). *Beginning to unlock the black box in the HR firm performance relationship: The impact of HR practices on employee attitudes and employee outcomes*. CAHRS Working Paper #01-12, Cornell University School of Industrial and Labor Relations, Center for Advanced Human Resource Studies, Ithaca.
7. Golding, N. (2010). Strategic Human Resource Management. In J. Beardwell & T. Claydon (Eds.), *Human Resource Management—a contemporary approach* (6th ed., pp. 30–82). New York, USA: Person Financial Times / Prentice Hall.
8. Guerry, M. (2011). Hidden heterogeneity in manpower systems: A Markov-switching model approach. *European Journal of Operational Research*, *210*(2011), 106–113.
9. Huselid, M. A., & Becker, B. E. (2011). Bridging micro and macro domains: Workforce differentiation and strategic Human Resource Management. *Journal of Management, 37*(2), pp. 421–428.
10. Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.
11. Liao, H., & Toya, K. (2009). Do They see eye to eye? Management and employee perspectives of high-performance work systems and influence processes on service quality. *Journal of Applied Psychology, 94*(2), pp. 371–391.
12. Macal, C. M., & North, M. J. (2006). Tutorial on agent-based modeling and simulation part 2: How to model with agents. In L. F. Perrone, et al. (Eds.), *Proceedings of the 2006 Winter Simulation Conference* (pp. 73–83). MD, USA: Baltimore.
13. Meixner, H.-E. (1987). *Personalstrukturplanung Leistung und Motivation durch Beförderungs- und Verwendungskonzepte*. Part 1: Ursachen und Folgen verbauter Berufs- und Karrierewege, Carl Heymanns Verlag, Cologne, Germany.
14. Morecroft, J. (2007). *Strategic modelling and business dynamics—a feedback systems approach*. Chichester: Wiley.
15. North, M. J., & Macal, C. M. (2007). *Managing business complexity—discovering strategic solutions with agent-based modeling and simulation*. New York, USA: Oxford University Press.
16. Purcell, J., & Kinnie, N. (2007). HRM and business performance. In P. Boxall, J. Purcell, & P. Wright (Eds.), *The Oxford handbook of Human Resource Management* (pp. 533–551). New York, NY, USA: Oxford University Press.
17. Scholl, H. J. (2001). *Agent-based and system dynamics modeling: A call for cross study and joint research*. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences, Wailea Maui, Hawaii, January 3–6, 2001.
18. Sterman, J. D. (2000). *Business dynamics—systems thinking and modeling for a complex world*. Boston: McGraw-Hill.
19. Wang, J. (2005). *A review of operations research applications in workforce planning and potential modelling of military training*. DSTO-TR-1688, Defence Science and Technology Organisation (DSTO), Systems Sciences Laboratory, Edinburgh, Australia.
20. Wright, P. M., & Boswell, W. R. (2002). *Desegregating HRM: A review and synthesis of micro and macro Human Resource Management Research*. CAHRS Working Paper #02-11, Cornell University School of Industrial and Labor Relations, Center for Advanced Human Resource Studies, Ithaca, NY, USA.

Re-Optimization of Rolling Stock Rotations

Ralf Borndörfer, Julika Mehrgardt, Markus Reuther, Thomas Schlechte and Kerstin Waas

Abstract The Rolling Stock Rotation Problem is to schedule rail vehicles in order to cover timetabled trips by a cost optimal set of vehicle rotations. The problem integrates several facets of railway optimization, such as vehicle composition, maintenance constraints, and regularity aspects. In industrial applications existing vehicle rotations often have to be re-optimized to deal with timetable changes or construction sites. We present an integrated modeling and algorithmic approach to this task as well as computational results for industrial problem instances of DB Fernverkehr AG.

1 Introduction

Rolling stock, i.e., rail vehicles, is the most expensive and limited asset of a railway company and must therefore be used efficiently. The *Rolling Stock Rotation Problem* (RSRP) deals with the cost minimal implementation of a railway timetable by constructing *rolling stock rotations* to operate passenger trips by rail vehicles. The RSRP integrates several operational requirements like *vehicle composition rules*, *maintenance constraints*, *infrastructure capacity constraints*, and *regularity requirements*.

T. Schlechte (✉) · R. Borndörfer · J. Mehrgardt · M. Reuther
Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany
e-mail: schlechte@zib.de

R. Borndörfer
e-mail: borndoerfer@zib.de

J. Mehrgardt
e-mail: mehrgardt@zib.de

M. Reuther
e-mail: reuther@zib.de

K. Waas
DB Fernverkehr AG, Frankfurt, Germany

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_8,

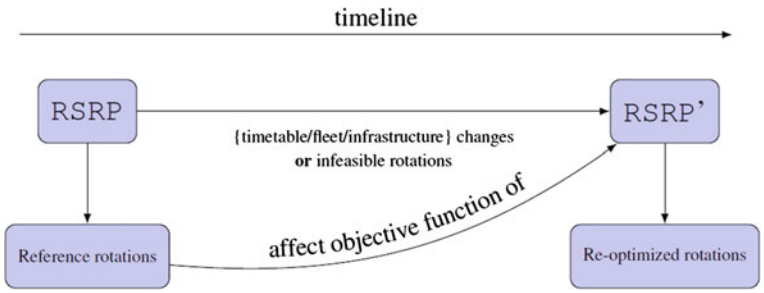


Fig. 1 Concept of re-optimization for the RSRP

A detailed problem description, a mixed integer programming formulation, and an algorithm to solve this problem in an integrated manner is described in detail in [3].

In this paper, we discuss one of the most important and challenging industrial applications of the RSRP, namely, *re-optimization*.

A re-optimization high-level concept for the RSRP is illustrated in Fig. 1. At some point in time a railway undertaking has to tackle an instance of the RSRP and constructs a solution (see boxes RSRP and Reference rotations). At another point in time this problem changes, such that the existing *reference rotation plan* can no longer be operated. Thus, a new problem RSRP has to be solved. The most important difference to the previous planning step is that much of the reference rotation plan was already implemented: Crew was scheduled for vehicle operations and maintenance tasks, capacity consumption of parking areas was reserved, and most important in a segregated railway system, e.g., in Europe and Germany: train paths were already allocated for the deadhead trips. Therefore a major goal is to change as little as possible in comparison to the reference rotation plan.

A literature overview on re-optimization can be found in [1, 2].

Re-optimization problems come up very often at a railway company. There are various causes that can lead to a situation where the implemented rotation plan becomes infeasible in an unexpected manner. Predictable and unpredictable construction sites are main causes. Fleet changes due to disruptions of operations or technical constraints, e.g., different maintenance constraints, modified speed limits for rolling stock vehicles, or changed infrastructure capacity, also ask for a modification of the vehicle rotation plans.

Depending on how large and how long the changes and their consequences are, re-optimization is required either by the dispatchers or in sufficiently lasting cases by the tactical and strategical divisions of the railway companies. In the latter case the problem is considered as a cyclic planning problem as it is introduced in [3].

The paper contributes an adaptation of the generic mixed integer programming approach presented in [3] to re-optimize rolling stock rotations. We show how to incorporate detailed re-optimization requirements into a hypergraph based formulation for rolling stock optimization by simply defining an appropriate objective function.

This paper is organized as follows. Section 2 defines the problem including an overview of our hypergraph based formulation. In Sect. 3 we introduce an objective modification procedure for the re-optimization case. Computational results in Sect. 4 show that our model and algorithm produce high quality and implementable results even for complicated re-optimization settings. Rotation planners of Deutsche Bahn validated the resulting rolling stock rotations from a detailed technical and operational point of view. It turned out that our configuration of the objective function described in Sect. 3 is sufficient and very precise for all re-optimization instances we got.

2 The Rolling Stock Rotation Problem

In this section we provide an overview on the hypergraph based rolling stock optimization model proposed in our previous paper [3]. We focus here on the main modeling ideas. For technical details including the treatment of maintenance and capacity constraints see [3]. The extension of the following problem description and model to include maintenance constraints is straight forward and does not affect the content or the contribution of the paper. Nevertheless, in our computational study we provide results for instances with maintenance constraints.

We consider a cyclic planning horizon of one *standard week*. The set of timetabled passenger trips is denoted by T. Let V be a set of *nodes* representing departures and arrivals of vehicles operating passenger trips of T, let $A \subseteq V \times V$ be a set of directed standard arcs, and $H \subseteq 2^A$ a set of *hyperarcs*. Thus, a hyperarc $h \in H$ is a set of standard arcs. The RSRP *hypergraph* is denoted by $G = (V, A, H)$. The hyperarc $h \in H$ *covers* $t \in T$, if each standard arc $a \in h$ represents an arc between the departure and arrival of t. We define the set of all hyperarcs that cover $t \in T$ by $H(t) \subseteq H$. By defining hyperarcs appropriately vehicle composition rules and regularity aspects can be directly handled by our model.

The RSRP is to find a cost minimal set of hyperarcs $H_0 \subseteq H$ such that each timetabled trip $t \in T$ is covered by exactly one hyperarc $h \in H_0$ and $\bigcup_{h \in H_0} a$ is a set of *rotations*, i.e., a set packing of cycles (each node is covered at most one time).

We define sets of incoming and outgoing hyperarcs of $v \in V$ in the RSRP hypergraph G as $H(v)^{in} := \{h \in H \mid \exists a \in h : a = (u, v)\}$ and $H(v)^{out} := \{h \in H \mid \exists a \in h : a = (v, w)\}$, respectively. By using a binary decision variable for each hyperarc, the RSRP can be stated as a mixed integer program as follows:

$$\min \sum_{h \in H} \mathbf{c}_h x_h, \tag{MP}$$

$$\sum_{h \in H(t)} x_h = 1 \qquad \forall t \in T, \tag{1}$$

$$\sum_{h \in H(v)^{in}} x_h = \sum_{h \in H(v)^{out}} x_h \quad \forall v \in V, \tag{2}$$

$$x \in \{0,1\}^{|H|} \tag{3}$$

The objective function of model (MP) minimizes the total cost of operating a timetable. For each trip $t \in T$ the covering constraints (1) assign exactly one hyperarc of $H(t)$ to t. The equalities (2) are flow conservation constraints for each node $v \in V$ that imply the set of cycles in the arc set A. Finally, constraints (3) state the integrality constraints for our decision variables.

3 Re-Optimization

The major re-optimization requirement for the RSRP is to change as little as possible in the reference rotation plan. We argue that this requirement can be handled by defining a suitable objective function based on the reference rotation plan.

$$\mathbf{c} : H \mapsto \mathbb{Q}_+ \; : \; \mathbf{c}(h) := \left\langle \begin{pmatrix} \mathbf{c}_1(h) \\ \mathbf{c}_2(h) \\ \mathbf{c}_3(h) \\ \mathbf{c}_4(h) \\ \mathbf{c}_5(h) \\ \mathbf{c}_6(h) \\ \mathbf{c}_7(h) \\ \mathbf{c}_8(h) \\ \mathbf{c}_9(h) \end{pmatrix}, \begin{pmatrix} p_1(h) \\ p_2(h) \\ p_3(h) \\ p_4(h) \\ p_5(h) \\ p_6(h) \\ p_7(h) \\ p_8(h) \\ p_9(h) \end{pmatrix} \right\rangle \begin{matrix} \ldots \textit{ connection deviations} \\ \ldots \textit{ composition deviations} \\ \ldots \textit{ rotation deviations} \\ \ldots \textit{ service deviations} \\ \ldots \text{ vehicles} \\ \ldots \text{ services} \\ \ldots \text{ deadhead distance} \\ \ldots \text{ regularity} \\ \ldots \text{ couplings} \end{matrix} \tag{4}$$

Definition (4) illustrates our approach. Our objective function is the sum of the re-optimization cost $\sum_{i=1}^{4} \mathbf{c}_i p_i$ and the original objective function $\sum_{i=5}^{9} \mathbf{c}_i p_i$. We propose to compute the parts of the re-optimization objective as a sum of costs depending on individual hyperarcs.

Let $h \in H$ be a hyperarc. In a first step we reinterpret h in the reference rotations, i.e., we search the timetabled trips that are connected or covered by h in the reference rotation plan, if they still exist. The reinterpretation procedure is very precise as a node in our hypergraph has the following attributes w.r.t. the vehicle traversing the node: position in a composition, orientation w.r.t. driving direction, fleet type, and rotation (i.e., cycle) of a vehicle.

In a second step we compute a property $p_i(h) \in \mathbb{N}$ for $i = 1, \ldots, 4$ for $h \in H$ that states the number of differences of h w.r.t. the reference rotations. Examples for such differences are:

Table 1 Key numbers of re-optimization scenarios

Instance	Trips	Compositions	Fleets	Maintenances	$\|V\|$	$\|H\|$
RSRP_11	104	2	1	0	486	186,130
RSRP_12	104	2	1	1	486	192,612
RSRP_13	104	2	1	2	486	198,758
RSRP_21	805	2	2	0	9,810	15,770,498
RSRP_22	805	2	2	2	9,810	18,768,740
RSRP_31	788	2	2	0	7,776	11,727,856
RSRP_32	788	2	2	2	7,776	14,019,208
RSRP_41	789	10	4	0	16,790	42,764,116
RSRP_42	789	10	4	4	16,790	54,640,466

- Let $h \in H$ a hyperarc connecting the timetabled trips t_1 and t_2. If t_1 and t_2 exist in the reference rotations and both trips are not connected there, we set $p_1(h) = |h|$. In all other cases we set $p_1(h) = 0$.
- If h covers trip t that exists in the reference rotations and is operated by a different vehicle composition than h, we set $p_2(h) \geq 1$, otherwise $p_2(h) = 0$. The exact numeric number depends on $|h|$, how these vehicles are oriented, which fleets are used etc.
- If h implies that t is operated in a different rotation we set $p_3(h) = |h|$, otherwise $p_3(h) = 0$.
- If h implies a different maintenance service before or after a timetabled trip $p_4(h) = 1$, otherwise $p_4(h) = 0$.

Solutions of re-optimization instances often have the characteristic that major parts of the reference rotations are not changed but some small parts have to be modified. In some cases, however, new timetabled trips have to be incorporated into the reference rotation plans. To handle this case we also have to consider properties of the original objective function $\sum_{i=5}^{9} \mathbf{c}_i\, p_i$ for re-optimization instances, i.e., costs for vehicles consumed by a hyperarc, costs for maintenance services, costs for deadhead distances, cost for irregularities, and costs for coupling activities. Finally all of these individual properties are multiplied by individual cost parameters $\mathbf{c}_i$, $i = 1, \ldots, 9$ that can be adjusted to the requirements of industrial use cases.

In this way we are able to handle a lot of technical re-optimization details simply by changing objective coefficients. As already mentioned, we were able to instantiate all technical re-optimization scenarios we got so far by this simple objective configuration procedure, i.e., by penalizing *local* deviations w.r.t. the reference rotations. This makes it possible to apply the general model and algorithm presented in [3] to solve re-optimization instances.

Table 2 Key numbers of re-optimization results with ROTOR 2.0 and CPLEX 12.5

Instance	Vehicles	Dev. heads	Dev. configurations	Dev. fleets	Dev. orientations	Gap (%)	hh:mm:ss
RSRP_11	9	0	0	0	0	0.00	00:03:13
RSRP_12	9	0	0	0	0	0.09	00:04:05
RSRP_13	9	1	0	0	0	1.65	00:05:35
RSRP_21	55	3	1	1	0	0.00	00:09:27
RSRP_22	55	1	0	0	0	0.15	03:34:00
RSRP_31	55	29	2	2	0	0.00	00:07:59
RSRP_32	55	30	2	2	0	0.28	01:17:17
RSRP_41	61	39	7	45	23	0.32	00:49:36
RSRP_42	59	40	7	42	17	0.91	02:48:20

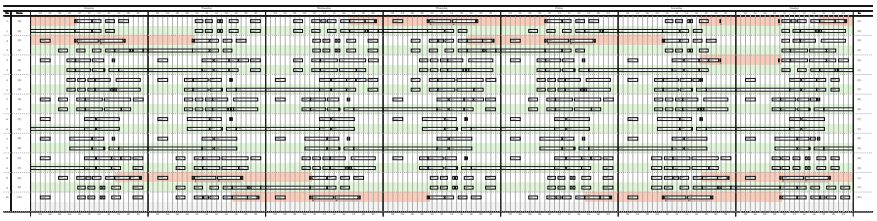


Fig. 2 Comparison of reference and re-optimized rolling stock rotations

4 Computational Results

We implemented our re-optimization model and algorithm in a computer program, called ROTOR 2.0. This implementation makes use of the commercial mixed integer programming solver CPLEX 12.5. ROTOR 2.0 is integrated in the IT environment of Deutsche Bahn. All our computations were performed on computers with an Intel(R) Xeon(R) CPU X5672 with 3.20 GHz, 12 MB cache, and 48 GB of RAM in multi thread mode with eight cores.

Table 1 lists the sizes of the instances, i.e., the number of *trips*, *compositions*, *fleets*, and *maintenance constraints*. In addition the total number of nodes ($|V|$) and hyperarcs ($|H|$) report about the size of the hypergraphs for the considered instances.

Furthermore, Table 2 provides re-optimization results. The second column reports on the number of used vehicles. The next four columns denote the number of deviations w.r.t. the reference solution introduced in Sect. 3. Finally, the last two columns show the proven worst case optimality gap and the total computation time.

The considered instances include scenarios where vehicles got broken, where the timetable was changed due to track sharing with other railway operators. And we also tackle instances where the fleet size increases, i.e., for the case when new vehicles are available and have to be integrated in the current operations. All scenarios were given by Deutsche Bahn Fernverkehr AG. Figure 2 shows a difference view of the

reference solution and the solution re-optimized with `ROTOR 2.0` in green. The rows alternate between the reference solution and the re-optimized solution. The red parts of the reference solution can never be reproduced because of timetable changes.

We conclude that re-optimization instances of Deutsche Bahn Fernverkehr AG for the RSRP can be handled in great detail. On the other hand huge parts of the reference rotation plans must not be changed: See column trips w.r.t. column $\sum p_1$ (sum of connection deviations) in Table 2 and Fig. 2. This combination directly results in short computation times, high quality solutions, and therefore a powerful tool for re-optimization of rolling stock rotations at Deutsche Bahn Fernverkehr AG.

References

1. Jespersen-Groth, J., Potthoff, D., Clausen, J., Huisman, D.,Kroon, L., Maróti, G., & Nielsen, M. (2009). Disruption management in passenger railway transportation. In A. RavindraK, M. RolfH, & Z. ChristosD (Eds.), *Robust and Online Large-Scale Optimization* (vol. 5868, pp. 399–421). Berlin: Springer.
2. Nielsen, L. K. (2011). Rolling stock rescheduling in passenger railways: Applications in short-term planning and in disruption management. *PhD Thesis*.
3. Reuther, M., Borndörfer, R., Schlechte, T., Waas, K., & Weider, S. (2013). Integrated optimization of rolling stock rotations for intercity railways. In *Proceedings of the* 5^{th} *ISROR (Rail Copenhagen 2013)*, Copenhagen, Denmark.

Branch-and-Price on the Split Delivery Vehicle Routing Problem with Time Windows and Alternative Delivery Periods

Heiko Breier and Timo Gossler

Abstract In this article we address the Split Delivery Vehicle Routing Problem with Time Windows and alternative Periods (SDVRPTWA). The consideration of multiple delivery periods per customer and the possibility of splitting deliveries across different periods makes it a relaxation of the well-known Vehicle Routing Problem with Time Windows and Split Deliveries (VRPTWSD). The problem is solved by a branch-and-price method. The opportunity for freight forwarders is to plan more efficient tours by exploiting alternative delivery periods. The contribution of this article is to prove the potential of this approach for cost savings and to demonstrate the decomposition of a SDVRPTWA in a demand focused master problem and period related pricing problems.

1 Introduction

Orders exceeding the vehicle capacity and customer induced time windows are common challenges in transport planning. As long as time windows are fix and the customer does not offer alternatives the problem can be solved with the well known VRPTWSD. In case that alternative non excluding delivery periods are offered, it would be possible to split the delivery across different periods. To our knowledge, this scenario has not been considered in the literature yet. Indeed, this methodology becomes more and more relevant as a growing number of companies introduces time window management systems requiring freight forwarders to book binding time windows. To face this requirement, the SDVRPTWA considers the possibility of serving

H. Breier (✉)
Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
e-mail: heiko.breier@kit.edu

T. Gossler
Accenture GmbH, 80807 Munich, Germany
e-mail: timo.gossler@accenture.com

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_9,

a customer either in one or in multiple periods. Thereby, it makes the selection of delivery periods subject of the optimization and offers the opportunity to plan tours with visits in alternative periods.

2 Related Problems

Research on the Vehicle Routing Problem with Time Windows (VRPTW) came up with [14] and [5] and with [9] as one of the most recent publications on solving the problem by column generation.

Splitting of deliveries was introduced by Dror and Trudeau [8]. A worst case analysis of split delivery problems was performed by Archetti et al. [1]. Archetti et al. [2] developed an efficient procedure to solve the Vehicle Routing Problem with Split Deliveries by column generation.

Frizzell and Giffin [10] were among the first authors examining the combination of split deliveries and time windows. The most relevant results to solve this problem by Branch-and-Price were presented by Gendreau et al. [11] and Desaulniers [6]. The two approaches differ mainly in regard to the decision, if quantities are considered within the pricing problem.

A good overview in general on the VRPTW can be found in [7] and in [3].

Cordeau [4] presented a Periodic Vehicle Routing Problem (PRVP). Here, each customer selects an individual combination of days. Visiting this customer is only possible at the selected combination of days. So, the possible combinations are input parameters for the program an not part of the optimization itself. Pirkwieser and Raidl [12] solved this problem by a column generation approach. In addition to a set-partioning formulation they introduce multiple pricing problems, one for each period. Their approach, anyhow, does not ensure integer solutions.

3 Problem Formulation

Given a set $\mathbb{V}$ which represents the nodes of the problem. Node 0 is the depot where all tours start and end. Set $\mathbb{I}$ represents the customers to deliver with $\mathbb{I} = \mathbb{V}\backslash 0$. $\mathbb{A}$ is the set of all arcs $(i, j) \in \mathbb{V} \times \mathbb{V}$. With $\mathbb{V}$ and $\mathbb{A}$ we can define a directed Graph $\mathbb{G} = (\mathbb{V}, \mathbb{A})$.

Additionally each customer defines which periods are valid to deliver goods. The set of periods is $\mathbb{P}$ and a customer i can choose any combination of periods $\mathbb{P}_i \subseteq \mathbb{P}$ to deliver him. Any delivery can have a split either within a time window in the same period or within time windows in different periods. These customer specific periods specify real alternatives. In an example this means: if $\mathbb{P} = [1, 2, 3, 4, 5]$ a customer i can choose $\mathbb{P}_i = [2, 3, 4]$ as alternative periods to receive deliveries. A possible

alternative could be to split up the delivery and to visit this customer in period 2 with the first part of the delivery and visiting him in period 4 to bring the last part of the delivery. This formulation is a relaxation of [4] and adds also the opportunity of split deliveries.

The problem can be written as a decomposed problem with a single Master Problem 3.1 and Sub Problems 3.2 for each period $p \in \mathbb{P}$.

All Sub Problems have to be solved separately for each period.

3.1 Master Problem

$\mathbb{R}$ is the set of available routes r in the Master Problem defined. $\mathbb{T}$ is the set of available delivery patterns t.

The decision variable is λ_{rt}^p which indicates the usage of a route $r \in \mathbb{R}$ in $p \in \mathbb{P}$ and a delivery pattern $t \in \mathbb{T}$. Each route has a length e_r. Constraints 2 ensure that each customer receives the complete demand d_i. ρ_{it} represents the delivery for customer i is visited in tour r with pattern t. The delivery pattern t is related to λ_{rt}^p. In constraints 3 we substitute each λ_{rt}^p with an variable y_{ij}^p, introduced by Desaulniers [6]. β_{ijr}^p is set to 1 if arc (i, j) is used in λ_{rt}^p, it is set to 0 otherwise. Variables y_{ij}^p will be used to perform branches (see Sect. 3.3). These variables represent how often an arc (i, j) is used in total or by period p as well as how often a customer i is visited in total or by period p. For all arcs between customers the usage is limited to $y_{ij}^p \leq 1$. For all arcs leaving/arriving at the depot usage of y_{ij}^p is unlimited and any value ≥ 0.4 is non-negative.

$$\min \left(\sum_{r \in \mathbb{R}, p \in \mathbb{P}, t \in \mathbb{T}} e_r \cdot \lambda_{rt}^p \right) \tag{1}$$

$$\text{s.t.} \sum_{r \in \mathbb{R}, p \in \mathbb{P}, t \in \mathbb{T}} \rho_{it} \cdot \lambda_{rt}^p \geq d_i \qquad \forall i \in \mathbb{I} \tag{2}$$

$$\sum_{r \in \mathbb{R}, t \in \mathbb{T}} \beta_{ijr}^p \lambda_{rt}^p = y_{ij}^p \quad \forall (i, j) \in \mathbb{V}, p \in \mathbb{P} \tag{3}$$

$$\lambda_{rt}^p \geq 0 \; \forall r \in \mathbb{R}, p \in \mathbb{P}, t \in \mathbb{T} \tag{4}$$

We start the solution procedure with a small set of columns in the Master Problem. These columns represent single trips from the depot to one customer and back to the depot. Also we add corresponding delivery patterns to those initial variables. With this first set of variables we start to generate additional variables λ_{rt}^p in a column generation process.

3.2 Pricing Problem

To get new valid routes for the Master Problem we solve an Elementary Shortest Path Problem with Ressource Constraints (ESPPRC) and include in the objective function the dual values of the Master Problem solved. We stop generating new λ_{rt}^{p} when no more columns with negative reduced cost can be found.

The objective of the problem is to find a route with lowest costs. α_{ij}^{p} and δ_i are the dual variables from the constraints 3 and 2 from the Master Problem presented. Variables x_{ij}^{p} decide if arc (i, j) is used in period p and are binary. Variables d_i^p decide which amount of the complete order d_i is delivered to customer i in period p and is integer. s_i^p and s_j^p, respectively, decide in which period p customers i, j are visited and are real values.

The distance between two nodes is c_{ij}. We have an unlimited set of identical vehicles T with capacity Q. Each customer $i \in \mathbb{I}$ has a demand $d_i > 0$ and a time window to deliver the goods. The customer specific interval to deliver starts at s_i^{start} and ends at s_i^{end}.

$$\min \left(\sum_{(i,j)\in\mathbb{V}} (c_{ij} + \alpha_{ij}^{p}) \cdot x_{ij}^{p} - \sum_{i\in\mathbb{I}} (\delta_i \cdot d_i^p) \right) \tag{5}$$

$$\text{s.t.} \sum_{i\in\mathbb{I}} x_{oj}^{p} = 1 \tag{6}$$

$$\sum_{j\in\mathbb{V}|i\neq j} (x_{ij}^{p} - x_{ji}^{p}) = 0 \qquad \forall i \in \mathbb{V} \tag{7}$$

$$\sum_{j\in\mathbb{I}} x_{i0}^{p} = 1 \tag{8}$$

$$\sum_{i\in\mathbb{I}} d_i^p \leq Q \tag{9}$$

$$\min(d_i; Q) \sum_{j\in\mathbb{V}|i\neq j} x_{ij}^{p} \geq d_i^p \qquad \forall i \in \mathbb{I} \tag{10}$$

$$s_i^{start} \leq s_i^p \leq s_i^{end} \qquad \forall i \in \mathbb{V}, p \in \mathbb{P}, t \in \mathbb{T} \tag{11}$$

$$s_i^p + b_{ij} - M(1 - x_{ijt}^{p}) \leq s_j^p \ \forall (i, j) \in \mathbb{V}, p \in \mathbb{P}, t \in \mathbb{T} \tag{12}$$

$$x_{ij}^{p} \in \{0; 1\} \qquad \forall (i, j) \in \mathbb{V} \tag{13}$$

$$d_i^p \in \{0, 1, \ldots\} \qquad \forall i \in \mathbb{I} \tag{14}$$

3.3 Branching Rules

When no more columns with negative reduced costs can be found, we test the solution of integrity. The solution is integral when the following branching rules deliver values which are integral. (i) branch on the vehicles used in total: $\sum_{r\in\mathbb{R},\, p\in\mathbb{P},\, t\in\mathbb{T}} \lambda_{rt}^{p}$ (ii) branch on the customers visited in total: $\sum_{p\in\mathbb{P},\, j\in\mathbb{V}} y_{ij}^{p}$ (iii) branch on the arcs used in total: $\sum_{p\in\mathbb{P}}(y_{ij}^{p} + y_{ji}^{p})$ (iv) branch on the vehicles used by period: $\sum_{r\in\mathbb{R},\, t\in\mathbb{T}} \lambda_{rt}^{p}$ (v) branch on the customers visited by period: $\sum_{j\in\mathbb{V}} y_{ij}^{p}$ (vi) branch on the arcs used by period: $(y_{ij}^{p} + y_{ji}^{p})$ (vii) branch on consecutive arcs: in [13] and [6] it is stated that also consecutive arcs are possible which generate non-integer solutions. In this case we branch on those consecutive arcs. However, those branches are only rarely necessary.

4 Results

The results are based on self generated instances for the problem. We created instances for up to 17 customers. The characteristics of the instances are specified by the available alternative periods and the demand of a customer.

Regarding the periods there are instances which allow (i) only a delivery at a single period p, (ii) at periods p or $p - 1$ and/or (ii) at periods p, $p - 1$ and/or $p - 2$. These instances are denoted as **s**ingle, **t**ight or **w**ide period instances.

Regarding the demand we created instances where (i) all demands do not exceed the capacity of a vehicle ($d_i \leq Q$), (ii) all demands exceed the capacity of a vehicle ($d_i > Q$) and (iii) the demands are a mix of the former ones. This is denoted by **d**eceeding, **e**xceeding or **m**ixed.

In an example instance 05dw is an instance with 5 customers, where demand does not exceed the capacity of a vehicle and three alternative periods for delivery are given.

Each of the instances was created for 4 different customer locations with different time windows, denoted by testset a–d.

The results are summarized in Table 1. The basis for the comparision of the savings is the related instance with a single period for delivery. For testinstance 07 dt (7 customers with a demand less or equal to the capacity of the vehicle and alternative periods for a delivery in p or $p - 1$ there are minimal savings of 5 % possible and maximal 15 % compared with instance 07 ds, where no alternative periods are allowed.

Further, the results show decreasing savings from deceeding instances to exceeding instances. The results show also that more alternative periods promise more savings. This is due to the fact that the forwarder has more alternatives to schedule an optimal transport plan.

Detailed results can be found in the appendix of this article.

Table 1 Minimal and maximal possible savings of the traveling distance

Demand	Periods		Customers						
			5 (%)	7 (%)	9 (%)	11 (%)	13 (%)	15 (%)	17 (%)
Deceeding	Tight	Min	0	5	8	2	7	15	0
		Max	13	15	20	20	19	22	10
	Wide	Min	6	11	9	2	14	21	1
		Max	24	20	23	20	32	32	23
Mixed	Tight	Min	0	4	6	3	6	10	/
		Max	9	12	15	16	14	17	/
	Wide	Min	4	5	10	3	7	10	/
		Max	14	15	22	19	20	26	/
Exceeding	Tight	Min	0	0	2	4	/	/	/
		Max	6	10	9	5	/	/	/
	Wide	Min	0	1	5	4	/	/	/
		Max	6	13	11	6	/	/	/

5 Conclusion

We presented in this article a Vehicle Routing Problem which includes Time Windows, Split Deliveries and alternative delivery periods. To the best of our knowledge this combination of restrictions/relaxations was never reviewed before.

Compared with problems without alternative delivery periods, using alternative delivery periods can achieve savings up to 32 % in our test instances. The idea of route scheduling with alternative periods is relevant in practice when freight forwarders have to book binding time windows at their destination

Appendix

In this section we present our solutions to the problem. We created instances with up to 17 customers

Table 2 Traveling distance for instances with 5–9 customers

Customers	Scenario	Testset a		Testset b		Testset c		Testset d	
		Distance	Savings (%)	Distance	Savings (%)	Distance	Savings (%)	Distance	Savings (%)
5	ds	216.56	0.00	229.65	0.00	158.21	0.00	253.83	0.00
	dt	187.68	13.34	229.65	0.00	158.21	0.00	223.16	12.08
	dw	164.77	23.91	193.59	15.70	148.9	5.88	222.72	12.26
	ms	366.5	0.00	340.59	0.00	231.15	0.00	364.75	0.00
	mt	335.32	8.51	340.59	0.00	231.15	0.00	333.64	8.53
	mw	314.71	14.13	304.53	10.59	221.84	4.03	333.64	8.53
	es	547.14	0.00	523.39	0.00	427.59	0.00	678.17	0.00
	et	546.4	0.14	523.39	0.00	427.59	0.00	637.73	5.96
	ew	523.82	4.26	505.27	3.46	425.9	0.40	637.73	5.96
7	ds	336.95	0.00	241.89	0.00	270.42	0.00	326.6	0.00
	dt	306.3	9.10	229.4	5.16	256.54	5.13	276.89	15.22
	dw	270.13	19.83	215.79	10.79	240.19	11.18	265.22	18.79
	ms	464.01	0.00	325.93	0.00	366.22	0.00	416.92	0.00
	mt	434.1	6.45	310.08	4.86	352.34	3.79	367.21	11.92
	mw	397.19	14.40	285.03	12.55	347.23	5.19	355.54	14.72
	es	814.76	0.00	667.92	0.00	717.65	0.00	745.4	0.00
	et	800.81	1.71	640.27	4.14	717.65	0.00	669.24	10.22
	ew	752.83	7.60	640.27	4.14	710.3	1.02	649.28	12.90
9	ds	319.16	0.00	393.64	0.00	409.67	0.00	342.89	0.00
	dt	270.56	15.23	358.53	8.92	377.58	7.83	275.25	19.73
	dw	253.51	20.57	358.53	8.92	353.85	13.63	264.19	22.95
	ms	607.82	0.00	622.14	0.00	713.35	0.00	544.61	0.00
	mt	557.53	8.27	587.03	5.64	668.01	6.36	461.83	15.20
	mw	511.05	15.92	553.14	11.09	641.91	10.01	424.8	22.00
	es	889.31	0.00	938.62	0.00	1,026.37	0.00	899.98	0.00
	et	807.35	9.22	918.22	2.17	979.8	4.54	849.77	5.58
	ew	792.1	10.93	890.47	5.13	976.34	4.87	824.04	8.44

Table 3 Traveling distance for instances with 11–17 customers

Customers	Scenario	Testset a		Testset b		Testset c		Testset d	
		Distance	Savings (%)	Distance	Savings (%)	Distance	Savings (%)	Distance	Savings (%)
11	ds	333.42	0.00	342.21	0.00	417.31	0.00	369.88	0.00
	dt	324.84	2.57	274.87	19.68	377.99	9.42	362.97	1.87
	dw	296.64	11.03	274.73	19.72	354.84	14.97	362.97	1.87
	ms	497.43	0.00	450.73	0.00	611.99	0.00	549.74	0.00
	mt	469.51	5.61	378.31	16.07	568.13	7.17	533.88	2.89
	mw	446.32	10.27	365.37	18.94	552.9	9.66	533.88	2.89
	es	940.98	0.00	863.31	0.00	1,163.74	0.00	962.35	0.00
	et	905.23	3.80		–	1,105.36	5.02		–
	ew	905.98	3.72		–	1,093.1	6.07		–
13	ds	588.4	0.00	500.96	0.00	587.13	0.00	471.3	0.00
	dt	492.46	16.31	438.46	12.48	545.1	7.16	380.24	19.32
	dw	401.11	31.83	438.46	12.48	505.71	13.87	340.11	27.84
	ms	978.39	0.00	920.02	0.00	968.78	0.00	863.41	0.00
	mt	863.25	11.77	861.29	6.38	877.6	9.41	746.12	13.58
	mw	789.41	19.32	858.37	6.70	818.2	15.54	723.11	16.25
	es	1,500.27	0.00	1,370.7	0.00	1,494.91	0.00	1,195.62	0.00
	et		–	1,345.31	1.85		–		–
	ew		–		–		–		–
15	ds	514.1	0.00	606.84	0.00	448.31	0.00	647.74	0.00
	dt	439.23	14.56	483	20.41	375.73	16.19	507.72	21.62
	dw	408.28	20.58	425.94	29.81	356.61	20.45	439.5	32.15
	ms	822.84	0.00	983.33	0.00	761.53	0.00	1,090.57	0.00
	mt	744.65	9.50	820.99	16.51	642.4	15.64	923.34	15.33
	mw		–	773.46	21.34	612.92	19.51	802.37	26.43
	es		–		–		–		–
	et		–		–		–		–
	ew		–		–		–		–
17	ds	511.73	0.00	564.42	0.00	536.06	0.00	534.07	0.00
	dt	466.71	8.80	508.67	9.88	481.06	10.26	530.56	0.66
	dw	454.45	11.19	433.85	23.13	467.72	12.75	527.24	1.28
	ms		–		–		–		–
	mt		–		–		–		–
	mw		–		–		–		–
	es		–		–		–		–
	et		–		–		–		–
	ew		–		–		–		–

References

1. Archetti, C., Savelsbergh, M., & Speranza, M. G. (2006). Worst-case analysis for split delivery vehicle routing problems. *Transportation Science*, *40*(2), 226–234.
2. Archetti, C., Bianchessi, N., & Speranza, M. G. (2011). A column generation approach for the split delivery vehicle routing problem. *Networks*, *58*(4), 241–254.
3. Baldacci, R., Mingozzi, A., & Roberti, R. (2012). Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints. *European Journal of Operational Research*, *218*(1), 1–6.
4. Cordeau, J.-F., Laporte, G., & Mercier, A. (2001). A unified tabu search heuristic for vehicle routing problems with time windows. *Journal of the Operational Research Society*, *52*(8), 928–936.
5. Desrosiers, J., Soumis, F., & Desrochers, M. (1984). Routing with time windows by column generation. *Networks*, *14*(4), 545–565.
6. Desaulniers, G. (2010). Branch-and-price-and-cut for the split-delivery vehicle routing problem with time windows. *Operations Research*, *58*(1), 179–192.
7. Desaulniers, G., Desrosiers, J., & Spoorendonk, S. (2010). The vehicle routing problem with time windows: State-of-the-art exact solution methods. In: J.J. Cochrane (Ed.), *Wiley encyclopedia of operations research and management science* (Vol. 8, pp. 5742–5749). Wiley. http://onlinelibrary.wiley.com/doi/10.1002/9780470400531.eorms1034/abstract.
8. Dror, M., & Trudeau, P. (1989). Savings by split delivery routing. *Transportation Science*, *23*(2), 141.
9. Feillet, D. (2010). A tutorial on column generation and branch-and-price for vehicle routing problems. 4OR: A Quarterly Journal of Operations Research.
10. Frizzell, P. W., & Giffin, J. W. (1995). The split delivery vehicle scheduling problem with time windows and grid network distances. *Computers and Operations Research*, *22*(6), 655–667.
11. Gendreau, M., Dejax, P., Feillet, D., & Gueguen, C. (2006). *Vehicle routing with time windows and split deliveries*. Laboratoire Informatique dAvignon, Technical report.
12. Pirkwieser, S., & Raidl, G. (2009). A column generation approach for the periodic vehicle routing problem with time windows. *Proceedings of the International Network Optimization Conference*.
13. Salani, M., & Vacca, I. (2009). *Branch and price for the vehicle routing problem with discrete split deliveries and time windows*. Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Technical report.
14. Solomon, M. M. (1983). Vehicle routing and scheduling with time windows constraints: Models and algorithms (PhD thesis, University of Pennsylvania).

Welfare Maximization of Autarkic Hybrid Energy Systems

Katja Breitmoser, Björn Geißler and Alexander Martin

Abstract Hybrid energy systems become a promising way for electrification of off-grid rural areas. We consider an autarkic mini-grid of households equipped with local solar panels, diesel generators and energy storage devices. Our aim is to find an energy distribution that maximizes the global welfare of the whole system. We present an MIQP model for the hybrid energy system optimization problem together with some remarks on computational results.

1 Introduction

Most research on stand-alone hybrid energy systems is focused on design and simulation rather than optimization of system control [2]. In this work, we discuss the latter problem in terms of finding a welfare-maximal power distribution of a hybrid energy system supplying a small community. We consider a decentralized system, where every household has its own solar panel, battery, and an optional diesel generator as backup device. The households are connected via a mini-grid to facilitate energy trading. The considered system is extended by an additional smart component that is able to defer the operation times of so-called smart devices on the consumer side to times with electricity excess.

2 Model Derivation

In this section, we derive the objective function together with the most important constraints of our model. We consider a set $\mathcal{N} = \{1, \ldots, N\}$ of households connected through a grid. The set of households equipped with a diesel generator is

K. Breitmoser (✉) · B. Geißler · A. Martin
Department of Mathematics, FAU Erlangen-Nürnberg, Cauerstraße 11,
91058 Erlangen, Germany
e-mail: Katja.Breitmoser@math.uni-erlangen.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_10,

denoted by $\mathcal{N}_D \subseteq \mathcal{N}$. Our planning horizon is subdivided into a set $\mathcal{I}$ of time intervals of length one hour. Further, we assume all households to have similar consumption and production possibilities and that the number of households is high enough such that no household has any market power.

2.1 Consumer Problem

We distinguish between profile loads and so called deferrable loads. For profile loads, the variables $x^{\text{pro}}_{n,i} \in \mathcal{X}^{\text{pro}}_{n,i}$ denote the aggregated demands of a household n in time interval i for all $n \in \mathcal{N}$ and $i \in \mathcal{I}$. The bounded sets $\mathcal{X}^{\text{pro}}_{n,i} \subset \mathbb{R}_{\geq 0}$ are the possible consumption quantities, which depend on historical load profiles but are assumed to be given herein. The demand functions $d_{n,i}(\pi) : [0, \overline{\pi}] \to \mathcal{X}^{\text{pro}}_{n,i}$ give the quantities of maximal utility for every price $\pi \in [0, \bar{\pi}]$ and are supposed to be strictly decreasing, bounded and continuous. The inverse demand, i.e., the marginal willingness to pay, is denoted by $p_{n,i}(x^{\text{pro}}_{n,i}) : \mathcal{X}^{\text{pro}}_{n,i} \to [0, \overline{\pi}]$. Then, the gross benefit $B_{n,i}$ of consumer n at time interval i is given by

$$B^{\text{pro}}_{n,i}(x^{\text{pro}}_{n,i}) = \int_0^{x^{\text{pro}}_{n,i}} p_{n,i}(z)\, dz. \tag{1}$$

In our model, we incorporate piecewise linear approximations of the demand functions by means of the so-called incremental method [4] and are thus able to give a closed-form expression of the gross benefit. Suppose that all consumers get the same price, then let $(\bar{x}_{n,i,h},\ \bar{\pi}_h)$, with $h \in \mathcal{H}$, denote the consumption bundles at the breakpoints of the approximation of the demand function $d_{n,i}$. Then, for each $n \in \mathcal{N}$ and $i \in \mathcal{I}$ the demand together with the marginal willingness to pay is modeled by

$$x^{\text{pro}}_{n,i} = \bar{x}_{n,i,1} + \sum_{h=1}^{|\mathcal{H}|-1} (\bar{x}_{n,i,h+1} - \bar{x}_{n,i,h})\, \delta_{i,h}, \tag{2a}$$

$$\pi_i = \bar{\pi}_1 + \sum_{h=1}^{|\mathcal{H}|-1} (\bar{\pi}_{h+1} - \bar{\pi}_h)\, \delta_{i,h}, \tag{2b}$$

$$1 \geq \delta_{i,1} \geq z_{i,1} \geq \cdots \geq z_{i,|\mathcal{H}|-2} \geq \delta_{i,|\mathcal{H}|-1} \geq 0, \tag{2c}$$

$$z_{i,h} \in \{0, 1\}, \quad \text{for } h = 1, \ldots, |\mathcal{H}| - 2. \tag{2d}$$

Thus, we obtain a unique representation of each $x^{\text{pro}}_{n,i} \in \mathcal{X}^{\text{pro}}_{n,i}$ in terms of the additional variables $\delta_i = (\delta_{i,1}, \ldots, \delta_{i,|\mathcal{H}|})$. A similar derivation of the gross benefit, according to a piecewise linear demand function, has been given in [5]. In our case,

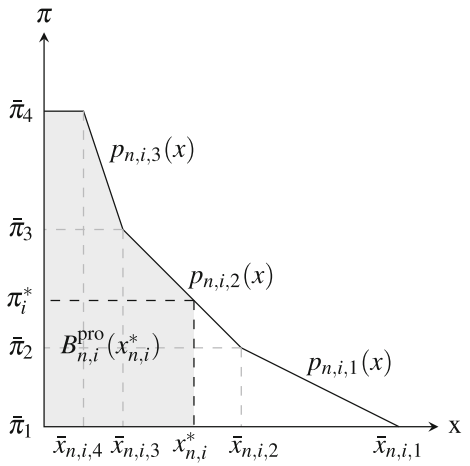


Fig. 1 Gross benefit of consumer n at time interval i for a consumption of $x_{n,i}^*$ and price π_i^*

the gross benefit, illustrated in Fig. 1, of consumer n at time interval i is given by

$$B_{n,i}^{\mathrm{pro}}(\delta_i) = (\bar{x}_{n,i,h+1} - \bar{x}_{n,i,h})\,(\delta_{i,h} - 1)\,\bar{\pi}_h + \frac{1}{2}(\bar{x}_{n,i,h+1} - \bar{x}_{n,i,h})\,(\bar{\pi}_{h+1} - \bar{\pi}_h)\,(\delta_{i,h}^2 - 1) + \bar{x}_{n,i,|\mathscr{H}|}\,\bar{\pi}_{|\mathscr{H}|}. \quad (3)$$

In contrast to profile loads, deferrable loads are not time dependent but rather have a specified run time and energy demand $X_{n,|\mathscr{I}|}^{\mathrm{def}} \in [\underline{X}_n^{\mathrm{def}}, \overline{X}_n^{\mathrm{def}}]$. The smart component must decide at which point in time, within a predefined time window $[\underline{\imath}_n, \bar{\imath}_n] \subseteq \mathscr{I}$, such a consumer is turned on. By $x_{n,i}^{\mathrm{def}}$ we denote the amount of power used for the smart devices of household n in time interval i for all $n \in \mathscr{N}$ and $i \in \mathscr{I}$. The set of possible consumption quantities is described in terms of binary variables $s_{n,i}^{\mathrm{on}}$ and s_n. The variables s_n indicate whether the energy demand of household n is met within the planning horizon, while $s_{n,i}^{\mathrm{on}} = 1$, if and only if voltage is fed to the deferrable loads of household n during time interval i, i.e.,

$$\underline{x}_n^{\mathrm{def}}\, s_{n,i}^{\mathrm{on}} \le x_{n,i}^{\mathrm{def}} \le \overline{x}_n^{\mathrm{def}}\, s_{n,i}^{\mathrm{on}}, \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}, \quad (4a)$$

$$X_{n,i}^{\mathrm{def}} = X_{n,i-1}^{\mathrm{def}} + x_{n,i}^{\mathrm{def}}, \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}, \quad (4b)$$

$$\underline{X}_n^{\mathrm{def}}\, s_n \le X_{n,|\mathscr{I}|}^{\mathrm{def}} \le \overline{X}_n^{\mathrm{def}}\, s_n, \qquad \forall n \in \mathscr{N}, \quad (4c)$$

$$s_{n,i}^{\mathrm{on}},\ s_n \in \{0, 1\}, \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}. \quad (4d)$$

The gross benefit $B_n^{\mathrm{def}}(X_{n,|\mathscr{I}|}^{\mathrm{def}})$ obtained from the satisfaction of demands from deferrable loads of each household $n \in \mathscr{N}$ is linear in the amount of energy consumed for these loads over the whole planning horizon if $X_{n,|\mathscr{I}|}^{\mathrm{def}} \in [\underline{X}_n^{\mathrm{def}}, \overline{X}_n^{\mathrm{def}}]$ and zero otherwise.

2.2 Producer Problem

The variables $y_{n,i}^{\text{sol}} \in \mathscr{X}_n^{\text{sol}} \subseteq \mathbb{R}_{\geq 0}$ denote the quantity of power produced by the solar panel of household n in time interval i for all $n \in \mathscr{N}$ and $i \in \mathscr{I}$. The corresponding cost functions $C_n^{\text{sol}}(y_{n,i}^{\text{sol}})$ are supposed to be affine.

Next, we introduce a variable $l_{n,i} \in [0, l_n^{\max}]$ for the battery charge level of household n in time interval i for all $n \in \mathscr{N}$ and $i \in \mathscr{I}$. The capacity of the battery in household n is denoted by $l_n^{\max}$. Additionally, we assume fixed initial battery charge levels $\bar{l}_{n,0}$ and lower bounds $l_n^{\min}$ for the battery charge levels at the end of the planning horizon to be given. The variable $l_{n,i}^{+} \in [0, l_{n,\max}^{+}]$ denotes the power used to charge the battery and $l_{n,i}^{-} \in [0, l_{n,\max}^{-}]$ denotes the power withdrawn from the battery of household n during time interval i. Discharging a battery is considered as a production facility with affine cost functions $C_n^{\text{bat}}(l_{n,i}^{-})$. To properly model battery charge levels we add the following constraints:

$$l_{n,i} = l_{n,i-1} + l_{n,i}^{+} - l_{n,i}^{-}, \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}, \tag{5a}$$

$$l_{n,i}^{+} \leq l_{n,\max}^{+}\, b_{n,i}, \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}, \tag{5b}$$

$$l_{n,i}^{-} \leq l_{n,\max}^{-}\, (1 - b_{n,i}), \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}, \tag{5c}$$

$$b_{n,i} \in \{0, 1\}, \qquad \forall n \in \mathscr{N}\ \forall i \in \mathscr{I}. \tag{5d}$$

Finally, for the diesel generators, we assume linear variable costs with non-sunk fixed costs $C_n^{\text{gen}}(s_{n,i}^{\text{gen}}, y_{n,i}^{\text{gen}}) = c_{n,v}^{\text{gen}}\, y_{n,i}^{\text{gen}} + c_{n,f}^{\text{gen}}\, s_{n,i}^{\text{gen}}$ as in [1]. Here, the binary variables $s_{n,i}^{\text{gen}} \in \{0, 1\}$ are used to decide whether the diesel generator of household $n \in \mathscr{N}_D$ is used for production during time interval i. The amount of generated power is denoted by $y_{n,i}^{\text{gen}} \in \mathscr{X}_n^{\text{gen}} \subseteq \mathbb{R}_{\geq 0}$. The non-sunk fixed costs and the variable costs of the generators are denoted by $c_{n,f}^{\text{gen}}, c_{n,v}^{\text{gen}} > 0$, respectively.

In general, the supply correspondence, i.e., the set of profit maximizing production quantities, without battery discharge, for household $n \in \mathscr{N}_D$ and given price π_i^* at time interval $i \in \mathscr{I}$ is given by

$$\begin{aligned} \arg\max \quad & \pi_i^* (y_{n,i}^{\text{sol}} + y_{n,i}^{\text{gen}}) - C_n^{\text{sol}}(y_{n,i}^{\text{sol}}) - C_n^{\text{gen}}(s_{n,i}^{\text{gen}}, y_{n,i}^{\text{gen}}) \\ \text{s.t.} \quad & y_{n,i}^{\text{sol}} \in \mathscr{X}_n^{\text{sol}}, \\ & \begin{pmatrix} y_{n,i}^{\text{gen}} \\ s_{n,i}^{\text{gen}} \end{pmatrix} \in \mathscr{X}_n^{\text{gen}}. \end{aligned} \tag{6}$$

Since any demand $d(\pi^*) \in (0, y_{\max}^{\text{gen}})$ would potentially lead to infeasibility of the overall model, using a supply function is not appropriate for cost structures with non-sunk fixed costs as illustrated in Fig. 2. In order to allow the whole set of profitable outputs instead, we incorporate the profit maximization problem underlying (6) into our model.

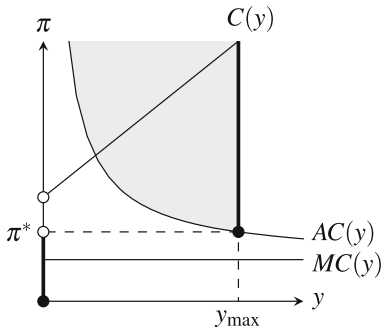


Fig. 2 Diesel generator costs: The corresponding production function is drawn with *thick lines*. Note, that for the price π^* the producer is indifferent between producing either $y_{\max}$ or nothing at all. The area, where production is profitable, is depicted in *gray*

2.3 Welfare-Maximal Power Distribution

The objective of our model is to maximize the global welfare of the whole community. That is the summed gross benefits of the loads minus the production costs:

$$\max \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} B^{\text{pro}}_{n,i}(\delta_i) + \sum_{n\in\mathcal{N}} B^{\text{def}}_{n}(X^{\text{def}}_{n,|\mathcal{H}|}) - \sum_{n\in\mathcal{N}_D} \sum_{i\in\mathcal{I}} C^{\text{gen}}_{n}(s^{\text{gen}}_{n,i}, y^{\text{gen}}_{n,i}) - \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} \left(C^{\text{sol}}_{n}(y^{\text{sol}}_{n,i}) + C^{\text{bat}}_{n}(l^{-}_{n,i}) \right), \tag{7}$$

subject to the constraints from above. Due to the gross benefit of the profile loads, the objective function is (convex) quadratic. Additionally, we have to add a clearing condition:

$$\sum_{n\in\mathcal{N}} \left(x^{\text{pro}}_{n,i} + x^{\text{def}}_{n,i} + l^{+}_{n,i} \right) = \sum_{n\in\mathcal{N}} \left(y^{\text{sol}}_{n,i} + l^{-}_{n,i} \right) + \sum_{n\in\mathcal{N}_D} y^{\text{gen}}_{n,i}, \qquad \forall i \in \mathcal{I}. \tag{8}$$

3 Results and Conclusions

We performed computational experiments on several test instances with 3–200 households, i.e., $|\mathcal{N}| \in \{3, 5, 15, 50, 100, 200\}$ and $|\mathcal{I}| \in \{24, 48, 72, 96\}$, where every fifth household is equipped with a diesel generator. We are thankful to the Siemens AG, for providing us with load and production profiles, see also [3]. All computations have been carried out with a time limit of 6 h on a computer with a 6-Core AMD Opteron 2435 processor with 2.6 GHz and 64 GB RAM. As MIQP-solver we used Cplex 12.5.0.0, which has been instructed to terminate, as soon as the relative optimality gap falls below 5 %. From the results we observed that only some of the largest instances, with 200 households and 96 time intervals, hit the time

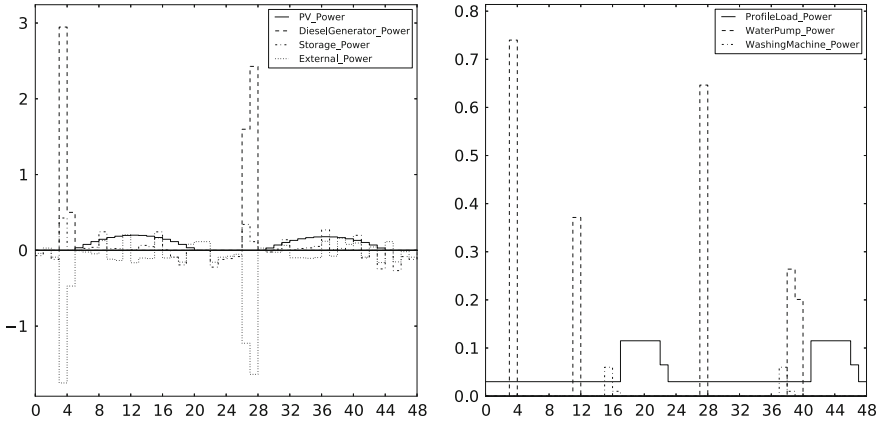


Fig. 3 Power distribution of the household with diesel generator

limit, whereas smaller instances, with up to 15 households, can reliably be solved within a few minutes.

Figure 3 illustrates the results for the single household, which is equipped with a diesel generator, from the solution of an instance with $|\mathcal{N}| = 3$ and $|\mathcal{I}| = 48$. To meet all demands, the power generated by the solar panels is not sufficient. Thus, the diesel generator runs once a day.

The power distributions of the two other households are basically similar, i.e., the deferrable loads are operated by the diesel generator or the excess power from the solar panel. The prioritization of the loads depends on the given demand and gross benefit functions. Besides, these functions ensure that every household receives as much energy as it is willing to pay.

Until now, our model does not account for compensation payments for households producing for other ones. For instance, the Shapley Value [6] defines fair and unique compensation payments for every participant. Consequently, future research could focus on an incorporation of the Shapley Value into our model. Beyond, polyhedral studies could help to close the gaps earlier and speed up computation.

References

1. Barley, C. D., & Winn, C. B. (1996). Optimal dispatch strategy in remote hybrid power systems. *Solar Energy*, *58*(4), 165–179.
2. Bernal-Agustń, J. L., & Dufo-López, R. (2009). Simulation and optimization of stand-alone hybrid renewable energy systems. *Renewable and Sustainable Energy Reviews*, *13*(8), 2111–2118.
3. Held, H., Klein, W., & Majewski, K., et al. (2012). Softgrid: A green field approach of future smart grid. In *Proceedings of the 2nd International Conference on Smart Grids and Green IT Systems*.

4. Markowitz, H. M., & Manne, A. S. (1957). On the solution of discrete programming problems. *Econometrica*, *25*, 84–110.
5. Martin, A., Müller, J. C., & Pokutta, S. (2012). Linear clearing prices in non-convex european dayahead electricity markets. Arxiv, preprint arXiv:1203.4177.
6. Shapley, L. S. (1953). Value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games II. Annals of Mathematics Study: Princeton University Press.*

Determining Optimal Discount Policies for a Supplier in B2B Relationships

Viktoryia Buhayenko and Erik van Eikenhorst

Abstract This research studies which discounts a supplier needs to offer to give incentive to his customers to change their order patterns in a way that minimizes the supplier's total cost. Savings for the supplier arise from reduction of set up and inventory cost. Customers also profit from this since the total discount offered is greater than their total cost increase. This research assumes zero or low price elasticity of the demand, thus lower prices do not result in greater total demand, they only affect when orders will be placed. A heuristic solution is given by separating the problem into when orders should be placed and how much discount should be offered to make this order pattern the cheapest for the buyer and includes three steps.

1 Introduction

Many articles in the field of operations management have analyzed ordering decisions while quantity discounts are in place. The problem of when and how much discount to offer is a problem that has received less attention, although it is of equally great practical importance as how to act on a given discount.

The problem of an optimal quantity discount schedule has first been analysed from supplier's perspective by Monahan [6]. He examines the situation with a sole customer and assumes that customer's demand is independent of discounts. An all-unit discount schedule and lot-for-lot policy are assumed. His model has been generalized and improved by Lee and Rosenblatt [5] who add a constraint on the amount of discount offered and drop the assumption of a lot-for-lot policy. The problem is

V. Buhayenko (✉)
Aarhus University, Fuglesangs Alle 4, 8210 Aarhus V, Denmark
e-mail: vbuhayenko@econ.au.dk

E. van Eikenhorst
Molde University College, Britvegen 2, 6410 Molde, Norway
e-mail: erik.van.eikenhorst@stud.himolde.no

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_11,

further developed in the article by Buscher and Lindner [1] who discover that the proposed algorithm can end up with an infeasible solution and modify it to avoid infeasibility. Rosenblatt and Lee [7] introduce the inventory holding costs of the supplier into the model and assume a linear discount schedule which results in a more equitable distribution of benefits. They also assume that the buyer tries to optimize his own objective function and doesn't always change his order size according to the supplier's desire.

The research is continued by Lal and Staelin [4] who extend their research to multiple buyer groups of different sizes. Chakravaty and Martin [2] also develop an algorithm for homogeneously grouped buyers. The problem is further analyzed assuming both all-units and incremental discounts with multiple break-points by Chen and Robinson [3] who research the situation when customers are heterogeneous only according to their demand. They use Pareto-type curve to describe the heterogeneity. The same discount policy is offered to all the customers. In all the papers mentioned above EOQ assumptions are used.

This research differs from the approaches stated above in the following way:

- dynamic demand and finite time horizon;
- a number of heterogeneous customers are considered, who are different not only in their demand but in their holding and order costs;
- discounts are different for every single customer;
- discounts can vary from period to period.

2 Problem Description

This research deals with the question of which discounts a supplier needs to offer to a set of heterogeneous to maximize his profit.

The supplier has a possibility to regulate the demand using discounts—to increase the demand in periods when the product is produced and to lower the demand in periods with no production of the product. Savings for the supplier arise from reduction of set up and inventory cost, but the buyer gets extra inventory cost. The discount offered by the supplier should be large enough to make the buyer order anyway at the period wanted by the supplier.

The major assumptions of this study are the following:

- Monopolistic situation of the supplier or very high barriers for switching suppliers so that price levels of other producers need not be considered here.
- Perfect information about the demand and cost of each buyer.
- Buyers and the supplier have no capital or warehouse restrictions.
- Buyers are considered to be rational in their reaction to the discount by choosing the lowest cost option available always.
- Buyers have full information in advance about future discounts.
- Simple discount in the form of a single price reduction.
- Single item case.

In the researched case, the total demand remains the same; there is only a question of when this demand is ordered and produced. This can appear, for example, when buyers are heavy equipment manufacturers and have stable demand for spare parts which are only a minor component of the final product [4].

3 Methodology and Solution Procedure

Exact solutions to the problem are very hard to achieve and would require an exponential amount of binary variables, representing each possible order schedule, for each customer.

Therefore, a heuristic solution approach has been developed which involves a separation between the problem when production and orders should take place, and the amount of discount that has to be offered to each costumer in each period to make them order at the periods indicated.

The following parameters are used while defining the algorithm:

d_{it} demand for every customer $i (i = 1, \ldots, n)$ in every period t. Demand of the supplier is the summation of his customers' orders in that period;
s_i fixed order processing/set up costs for every customer i and the supplier $i = 0$;
h_{it} carrying charge for each customer i and the supplier in each period t;
c_i initial costs for every customer i and the supplier $i = 0$;

We operate with the following decision variables:

H_{it} inventory for every customer i and supplier $i = 0$ in every period t;
S_{it} binary variable for every customer i and supplier $i = 0$ in every period t, 1, when the order/set up is made, 0 otherwise;
P_i total amount of compensation for every customer i;
Q_{it} order quantity for the customer i and the supplier in every period t.

The solution procedure includes the following three steps:

STEP 0. The problem for each customer and the supplier is solved using the Wagner-Whitin algorithm [8]. Initial order and production patterns are obtained on this stage. We also get initial costs c_i for every customer i.

STEP 1. Supplier's costs and compensations offered by the supplier to the customers are minimized. The compensations are a lump sum that would compensate the customer for ordering at the periods requested and represent increase in customers' costs. Minimizing compensations we in the end minimize customers' costs. Hence, the model offered below modifies initial order and production patterns ensuring both supplier and customers' cost minimization.

Objective (1) is minimized for the supplier:

$$\text{Minimize} \qquad \sum_{t=1}^{m} (h_{0t} H_{0t} + s_0 S_{0t}) + \sum_{i=1}^{n} P_i, \tag{1}$$

subject to

$$\sum_{t=1}^{m}(h_{it}H_{it} + s_iS_{it}) - P_i \leq c_i, \quad \text{for } \forall i > 0, \tag{2}$$

Constraint (2) ensures that there is no cost increase for any customer i.

There are two additional constraints (3) and (4) for the customers:

$$\sum_{u\geq t}^{m} d_{iu}S_{it} - Q_{it} \geq 0, \quad \text{for } \forall i > 0, \forall t, \tag{3}$$

$$Q_{it} + H_{it-1} - H_{it} = d_{it}, \quad \text{for } \forall i > 0, \quad \forall t, \tag{4}$$

For the supplier we have constraints (5) and (6):

$$\sum_{k=1}^{n}\sum_{u\geq t}^{m} d_{ku}S_{0t} - Q_{0t} \geq 0, \quad \text{for } \forall t, \tag{5}$$

$$Q_{0t} + H_{0t-1} - H_{0t} = \sum_{k=1}^{n} Q_{kt}, \quad \text{for } \forall t, \tag{6}$$

Constraints (4) and (6) ensure the continuity of the flow.

Constraints (3) and (5) link binary variables S_{it} with continuous variables Q_{it} forcing S_{it} to take value 1, when $Q_{it} \geq 0$, and 0, when $Q_{it} = 0$.

STEP 2. Discounts offered to achieve the order patterns determined at the previous step are calculated now.

A new variable accounting for a discount for every customer i in period t is introduced. An equation is solved for every customer i for every period t providing that a discount is introduced in this period. The left-hand side represents the difference between inventory and purchasing costs of the initial and final order patterns where the unit price is unknown; the right-hand side is the increase in costs while changing these patterns.

This is however a heuristic solution, because the total discount that has to be offered to make it cheaper than all other order patterns is larger than the increase in cost which is the compensation assumed in Step 1.

4 Numerical Example

The data for the considered numerical example was generated randomly for the problem size of 5 customers and 20 periods. We assume initial inventory to be equal to 0 in this example.

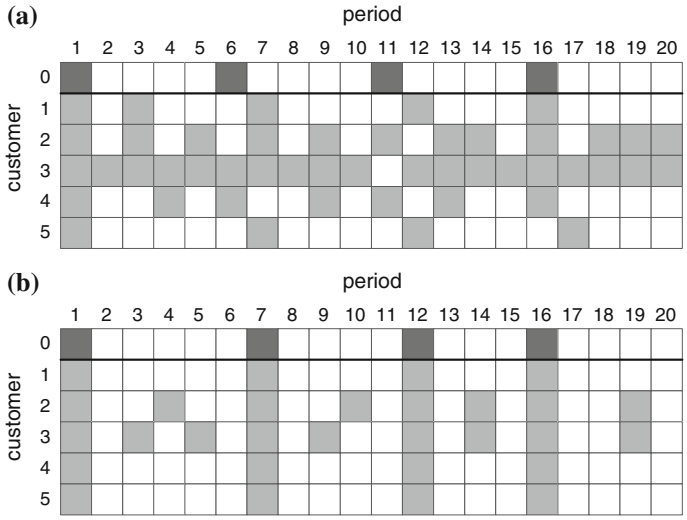


Fig. 1 Difference between the initial and final order/production patterns. **a** Initial order pattern. **b** Final order pattern

Figure 1 shows initial and final order and production patterns after implementation of the first two steps of the procedure described above. It displays the period in which the order/set up is done. The first row (customer 0) represents set up schedules of the supplier.

It can be noticed that in the initial pattern customers' orders rarely coincide with production periods. Customer 3 orders almost in every period. Customers 2 and 4 order very often as well. Despite rather low frequency of orders of customers 1 and 5 their order patterns are not synchronized with production patterns.

In the final pattern orders of customers 1, 4 and 5 totally match production periods. Customers 2 and 3 still order in-between production periods but their orders became significantly less frequent.

Finally at Step 2 the minimum discount needed in each order period is calculated so that the order pattern of Step 1 becomes the cheapest order pattern for this customer. It is assumed that the regular price is 100 per item. The resulting prices minus discounts for each customer are given in Table 1. Periods where no discount is introduced are omitted.

Result for the supplier is summarized in Table 2. Due to the implemented discount pricing schedule the supplier's profit was improved by 10,315.

Table 1 Discount schedule and total discount for every customer

	Total discount	Price									
		1	3	4	5	7	9	12	14	16	19
1	2,290.37	92.24									
2	7,286.66	95.08		95.67		90.42		99.90		86.55	99.45
3	11,107.55	96.92	98.34		97.70	95.42	89.37	99.06	97.03	86.02	95.13
4	13,908.61	82.20				86.98		97.60		98.78	
5	212.07									97.81	

Table 2 Result for the supplier

Original production and inventory cost	225,780
Cost reduction due to coordination of orders	45,120
Sales revenue lost because of discounts offered	34,805
Additional profit	10,315

5 Conclusions and Further Research

The supplier has a possibility to regulate the demand—to increase it in periods when the product is produced and to lower it in periods with no production of the product.

Savings for the supplier arise from reduction of inventory costs and sometimes reduction of set up costs.

The customer gets extra inventory costs which are compensated for with the discount offered by the supplier which should be large enough to make the customer order at the period wanted by the supplier. The customer makes a profit since the total discount is bigger than his increase in costs.

Further research will include transition from discount in form of a simple price reduction to all-units discount, extension of the problem to the multiple item case and introduction of a capacity constraint.

References

1. Busher, U., & Lindner, G. (2004). Ensuring feasibility in a generalized quantity discount pricing model to increase supplier's profits. *Journal of the Operational Research Society*, *55*, 667–670.
2. Chakravarty, A. K., & Martin, G. E. (1988). An optimal joint buyer-seller discount pricing model. *Computers and Operational Research*, *15*, 271–281.
3. Chen, R. R., & Robinson, L. W. (2012). Optimal multiple-breakpoint quantity discount schedules for customers with heterogenous demands: all-unit or incremental? *IIE Transactions*, *44*, 199–214.
4. Lal, R., & Staelin, R. (1984). An approach for developing an optimal discount pricing policy. *Management Science*, *30*(12), 1524–1539.
5. Lee, H. L., & Rosenblatt, M. J. (1986). A generalized quantity discount pricing model to increase supplier's profits. *Management Science*, *32*(9), 1177–1185.
6. Monahan, J. P. (1984). A quantity discount pricing model to increase vendor profits. *Management Science*, *30*(6), 720–726.

7. Rosenblatt, M. J., & Lee, H. L. (1985). Improving profitability with quantity discounts under fixed demand. *IIE Transactions*, *17*(4), 388–394.
8. Wagner, H., & Whitin, T. (1958). Dynamic version of the economic lot size model. *Management Science*, *5*(1), 89–96.

Exact and Compact Formulation of the Fixed-Destination Travelling Salesman Problem by Cycle Imposement Through Node Currents

Mernout Burger

Abstract The Travelling Salesman Problem (TSP) has been studied extensively for over half a century, but due to its property of being at the basis of many scheduling and routing problems it still attracts the attention of many research. One variation of the standard TSP is the multi-depot travelling salesman problem (MTSP) where the salesmen can start from and return to several distinct locations. This article focusses on the MTSP with the extra restriction that each salesman should return to his home depot, known as the fixed-destination MTSP. This problem (and its variations such as the multi-depot vehicle routing problem) is usually formulated using three-index binary variables, making the problem computationally expensive to solve. Here an alternative formulation is presented using two-index binary variables through the introduction of a limited amount of continuous variables to ensure the return of the salesmen to their home depots.

1 Introduction

The TSP has been a topic of research for over six decades [1], but it still attracts the attention of researchers due to its challenges and wide applicability. Many variations of the TSP have been introduced to model a real-world problem, such as the vehicle routing problem [12] and its many variations [8].

In this article we focus on the formulation of the TSP with multiple depots, were each salesman should return to his home depot. When considering scheduling and routing problems with multiple depots where each entity (e.g. a salesman or vehicle) should return to the home depot, we talk about fixed-destination problems [2]. Such problems are often formulated as a mixed-integer linear program (MILP) using a three-index formulation of binary variables

M. Burger (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: mernout@ocandor.com

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_12,

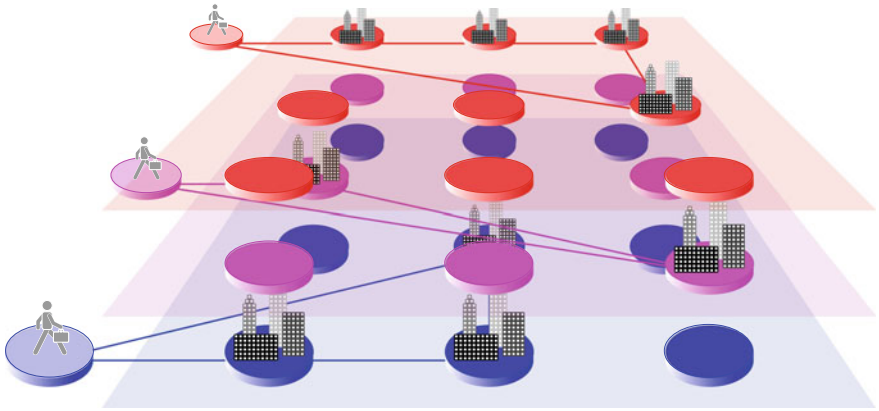


Fig. 1 A fixed-destination solution using a 3-index formulation for 3 depots and 9 cities

$$x_{ijk} = \begin{cases} 1 & \text{if location } i \text{ precedes location } j \text{ directly in a tour started at depot } k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

resulting in $\mathsf{C}^2\mathsf{D}$ binary variables for a problem with C cities and D depots. This problem can be depicted by a layered graph as shown in Fig. 1, where each depot has a copy of all city nodes in a separate layer. When solving a MILP using standard solvers, the computation time is largely dependent on the number of integer (binary) variables that are used to represent the problem. Therefore, it is beneficial to try to reduce the number of binary variables.

Recently an alternative formulation using two-index binary variables has been presented in [3] using a multi-commodity formulation. The set of depot nodes is cloned to create sink and source nodes for the commodity flows. Using D continuous variables (representing the commodities) at each of the $\mathsf{C} + 2\mathsf{D}$ locations (cities plus depots) it is ensured that a flow of commodities starting at a (source) depot will end at the associated copied (sink) depot, thereby ensuring fixed-destination solutions.

In this article an alternative two-index formulation is presented that requires a little less binary variables and significantly less continuous variables as compared to the multi-commodity formulation. There is no need to copy the depot nodes, and only one additional continuous variable per location is needed, resulting in an increase of $\mathsf{C} + \mathsf{D}$ continuous variables compared to the (non-fixed destination) TSP. These continuous variables can be seen as node currents, inspired by the subtour elimination constraints using node potentials that were introduced by Miller et al. [10]. This formulation has been used for solving scheduling problems for micro-ferries [5] and harvesters [6]. Here we will discus the method in detail for the basic MTSP to make readers aware of the possibility to use this formulation as the basis of other multi-depot scheduling and routing problems.

2 Fixed-Destination Travelling Salesman Problems

We will discuss the TSP with multiple depots, where each salesman should return to its home depot at the end of his tour. A novel formulation for this fixed-destination MTSP using two-index decision variables will be presented next.

2.1 Node Potentials and Currents

The inspiration of this approach comes from the *subtour elimination constraints* of Miller et al. [10] using *node potentials*. To avoid cycles in (a part of) a graph one can assign continuous variables to the nodes representing a potential in an electric circuit, and add constraints on their values to avoid subtours. We reckoned that if there are node potentials in a network, and the nodes are connected by arcs, there should also be arc currents flowing between the nodes. Since for a solution to the MTSP each node has exactly one incoming and one outgoing arc (see Fig. 2) this current can be seen as a property of the nodes (instead of the arcs). We will present a methodology that can be seen as the dual to the MTZ subtour elimination constraints; *cycle imposement constraints* using *node currents*.

2.2 Description of the Problem

Consider a problem with D depots and C cities with sets $\mathscr{D}$ and $\mathscr{C}$ defined as

$$\mathscr{D} = \{1, \ldots, \mathsf{D}\}, \quad \mathscr{C} = \{\mathsf{D}+1, \ldots, \mathsf{N}\}, \quad \mathscr{N} = \mathscr{D} \cup \mathscr{C}, \tag{2}$$

where $\mathsf{N} = \mathsf{D} + \mathsf{C}$ denotes the total number of locations represented by the set $\mathscr{N}$. This problem can be depicted by a graph with N nodes, where associated with each possible directed arc (i, j) we define a decision variable

$$x_{ij} = \begin{cases} 1 & \text{if there is a connection from node } i \text{ to node } j, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

resulting in a total of C^2 binary variables in the MILP formulation.

As shown in Fig. 2 the graph can be split into two subgraphs: the nodes in $\mathscr{D}$ are associated with the depots and the nodes in $\mathscr{C}$ are associated with the cities. From each of the depots we want one salesman[1] to travel towards a city (represented by an

[1] It is possible to formulate this problem with multiple salesmen per depot as well. To avoid distraction from the main purpose of this section the problem is kept as simple as possible.

arc from $\mathscr{D}$ to $\mathscr{C}$) and returning to his home depot at the end of the tour (represented by an arc from $\mathscr{C}$ to $\mathscr{D}$).

Although cycles in the set $\mathscr{C}$ must be avoided to obtain a correct solution, within the set $\mathscr{N}$ we want *exactly* D cycles; one associated with each of the depots in $\mathscr{D}$ (see Fig. 2). To obtain such a solution we introduce N continuous variables k_i that can be seen as the dual to the node potentials u_i; they can be considered node currents. To *impose* the existence of D cycles in the graph we give each depot node an unique value and propagate it along the path.

2.3 Formulation of the Problem

The fixed-destination MTSP can be formulated as the mixed-integer linear program[2]

$$\text{minimise} \quad \sum_{i\in\mathscr{N}} \sum_{j\in\mathscr{N}} \mathsf{c}_{ij} x_{ij} \tag{4a}$$

$$\text{subject to} \quad \sum_{j\in\mathscr{C}} x_{hj} = 1, \quad \sum_{i\in\mathscr{C}} x_{ih} = 1 \qquad \forall\, h \in \mathscr{N} \tag{4b}$$

$$u_i - u_j + \mathsf{N} x_{ij} \le \mathsf{N} - 1 \qquad \forall\, i, j \in \mathscr{C} \tag{4c}$$

$$k_d = d \qquad \forall\, d \in \mathscr{D} \tag{4d}$$

$$k_i - k_j + (\mathsf{D} - 1) x_{ij} \le \mathsf{D} - 1 \qquad \forall\, i, j \in \mathscr{N} \tag{4e}$$

$$x_{ij} \in \{0, 1\} \qquad \forall\, i, j \in \mathscr{N} \tag{4f}$$

where (4a) is the objective function representing the total travel distance, (4b) are the assignment constraints ensuring that each location is visited once and only once, and (4c) are the subtour elimination constraints. Using (4d) each variable k_d of the depot nodes is given a unique value, and (4e) propagate the value $k_i = d$ along cities i in the path of depot d. Note the strong resemblance to (4c); constraints (4c) might appear to be weaker versions of the subtour elimination constraints, but they actually impose the existence of D cycles (one for each depot) in the set $\mathscr{N}$ as explained next.

2.4 Node Current Propagation in Detail

In order to show that inequalities (4e) indeed enforce fixed-destination solutions (in combination with the assignment constraints (4b) and the subtour elimination constraints (4c)), we start by analysing the inequalities.

[2] To see the relation between node potentials and node currents more clearly the original subtour elimination constraints from [10] are used in (4c). For actual implementation these constraints can be made tighter using the formulations presented in [7, 11].

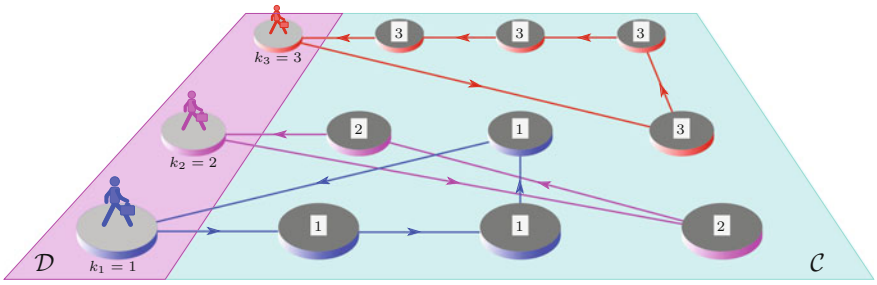


Fig. 2 Constraints (4e) ensure the existence of D cycles. This figure shows an example solution to the problem with $\mathsf{D} = 3$ depots and $\mathsf{C} = 9$ cities

When there is no direct path from location i to j we have $x_{ij} = 0$ and hence

$$k_i - k_j \leq \mathsf{D} - 1. \tag{5}$$

Since the cities will be associated to a depot with index number 1 to D, we expect the variable k_i to have a value in between 1 and D due to the equality constraints (4d). Therefore, inequality (5) is non-restrictive since $k_i \leq \mathsf{D}$ and $k_j \geq 1$. When $x_{ij} = 1$ it means that the path of a salesman goes from location i to j directly, and

$$k_i - k_j + \mathsf{D} - 1 \leq \mathsf{D} - 1 \quad \Leftrightarrow \quad k_i \leq k_j. \tag{6}$$

Hence, the value of k_j will be non-decreasing along the path. Since these inequalities should also hold for arcs (c, d) from a city c to depot d—and the value k_d of the depot is fixed by (4d)—a path that originates from depot d cannot return to a depot with a lower index number.

Now consider depot D. Since $k_i \leq k_j$ along each path we have $k_i \geq \mathsf{D}$ along the path originating from this depot. Due to the assignment constraints (4b) each node will have exactly one incoming arc and one outgoing arc, hence the path can only end in a depot node (otherwise there will be a city node with two incoming arcs). The only depot node d that can satisfy the constraint $k_d \geq k_c \geq \mathsf{D}$ is depot $d = \mathsf{D}$. Constraints (4d) and (4e) impose the existence of a cycle containing node D, and since $\mathsf{D} \leq k_c \leq k_d = \mathsf{D}$ we have $k_c = \mathsf{D}$ along the path of depot D.

Next consider depot node $\mathsf{D} - 1$. Along the path of this depot we have $k_c \geq \mathsf{D} - 1$, and since $k_d \geq k_c \geq \mathsf{D} - 1$ should hold when going from city c to depot d, the index number of the depot should be at least $\mathsf{D} - 1$. Since we know that depot D already has an incoming arc [and only one is allowed due to (4b)] the path started at depot $\mathsf{D} - 1$ can only return to depot $\mathsf{D} - 1$. Also $k_c = \mathsf{D} - 1$ along the path of depot $\mathsf{D} - 1$.

Continuing this reasoning one can see that each depot d has a path that returns to depot d, and $k_c = d$ along the path associated with depot d. Hence we have a solution with (at least) D cycles. Due to the subtour (cycle) elimination constraints

(4c) it is ensured that there are no cycles in $\mathscr{C}$; exactly D cycles exist in the graph, each associated with one of the depots. This resembles the solution to the fixed-destination MTSP, since each path returns to its home depot.

3 Conclusions

In this article we have demonstrated the use of node currents and cycle imposement constraints to formulate the fixed-destination travelling salesman problem as a mixed-integer linear program using two-index binary variables. The use of two-index formulations over three-index formulations results in shorter computation times and lower memory use when solving the problems using standard MILP solvers. Although specialised algorithms might outperform the standard MILP solvers, it is believed that the presented formulation might provide great benefits in solving variations of the multi-depot travelling salesman problem (such as the micro-ferry scheduling problem [5] and the multiple harvester routing problem [6]) for which specialised algorithms are (not yet) available.

References

1. Applegate, D. L., Bixby, R. E., Chvátal, V., & Cook, W. J. (2006). *The traveling salesman problem: a computational study*. Princeton: Princeton University Press.
2. Bektaş, T. (2006). The multiple traveling salesman problem: an overview of formulations and solution procedures. *Omega*, *34*(3), 209–219.
3. Bektaş, T. (2012). Formulations and Benders decomposition algorithms for multidepot salesmen problems with load balancing. *European Journal of Operational Research*, *216*(1), 83–93.
4. Burger, M., De Schutter, B., & Hellendoorn, J. (2012). Micro-ferry scheduling problem with time windows. In proceedings of American Control Conference (pp. 3998–4003)
5. Burger, M., & De Schutter, B. (2013). Energy-efficient transportation over flowing water. In proceedings of International Conference on Sensing and Control (ICNSC), pp. 226–231
6. Burger, M., Huiskamp, M., & Keviczky, T. (2013). Complete field coverage as a multiple harvester routing problem. In proceedings of Agriculture Control.
7. Desrochers, M., & Laporte, G. (1991). Improvements and extensions to the Miller-Tucker-Zemlin subtour elimination constraints. *Operations Research Letters*, *10*(1), 27–36.
8. Golden, G., Raghavan, S., & Wasil, E. (2008). *The vehicle routing problem: latest advances and new challenges*. Berlin: Springer.
9. Kara, I., & Bektaş, T. (2006). Integer linear programming formulations of multiple salesman problems and its variations. *European Journal of Operational Research*, *174*(3), 1449–1458.
10. Miller, C. E., Tucker, A. W., & Zemlin, R. A. (1960). Integer programming formulation of traveling salesman problems. *Journal of the ACM*, *7*(4), 326–329.
11. Sherali, H. D., & Driscoll, P. J. (2002). On tightening the relaxations of Miller-Tucker-Zemlin formulations for asymmetric traveling salesman problems. *Operations Research*, *50*(4), 656–669.
12. Toth, P., & Vigo, D. (2002). *The vehicle routing problem*. SIAM

0–1 Multiband Robust Optimization

Christina Büsing, Fabio D'Andreagiovanni and Annie Raymond

Abstract We provide an overview of new theoretical results that we obtained while further investigating *multiband robust optimization*, a new model for robust optimization that we recently proposed to tackle uncertainty in mixed-integer linear programming. This new model extends and refines the classical Γ-robustness model of Bertsimas and Sim and is particularly useful in the common case of arbitrary asymmetric distributions of the uncertainty. Here, we focus on uncertain 0–1 programs and we analyze their robust counterparts when the uncertainty is represented through a multiband set. Our investigations were inspired by the needs of our industrial partners in the research project ROBUKOM [2].

1 Introduction

Over the last years, professionals dealing with real-world optimization problems have increased their interest in embedding uncertainty in their decision process, showing particular attention to tractable *robust optimization* (RO) techniques. The goal of RO is to find an optimal solution that is deterministically protected against the worst coefficient deviations specified by an uncertainty set (we refer the reader to [3, 4]

C. Büsing
Department of Operations Research, RWTH Aachen University, Kackertstr. 7,
52072 Aachen, Germany
e-mail: buesing@or.rwth-aachen.de

F. D'Andreagiovanni (✉)
DFG Research Center MATHEON, Technical University Berlin, Straße des 17. Juni 135,
10623 Berlin, Germany
e-mail: d.andreagiovanni@zib.de

F. D'Andreagiovanni · A. Raymond
Department of Optimization, Zuse-Institut Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany
e-mail: raymond@zib.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_13,

for a comprehensive introduction to theory and applications of RO). Among the RO models proposed over the years, the Γ-robustness model of Bertsimas and Sim (Γ-Rob) [5] was a breakthrough in the development of tractable robust counterparts and has without doubt been the most successful and widely applied model. However, as pointed out by its authors, the assumptions at the basis of Γ-Rob may sensibly limit the possibility of modeling arbitrary-shaped distributions of the uncertainty that are commonly found in real-world problems, and lead to overconservative robust solutions (for a more detailed discussion of the limits of Γ-Rob, we refer the reader to [2, 6, 7]). Starting with the work [6], we have studied the possibility of refining Γ-Rob by exploiting a very simple operation: partitioning the single deviation band into multiple bands, each with its own parameters. This operation is at the basis of the general theoretical study that we have started to fill the gap of knowledge about the use of a multiband uncertainty set in RO.

2 Multiband Uncertainty

We consider a generic uncertain mixed-integer linear program (MILP):

$$\max \sum_{j \in J} c_j \, x_j \quad \text{s.t.} \quad \sum_{j \in J} a_{ij} \, x_j \leq b_i \, , \quad i \in I = \{1, \ldots, m\},$$
$$x_j \geq 0 \quad , \quad j \in J = \{1, \ldots, n\}, \quad x_j \in \mathbb{Z}^+, \quad j \in J^{\mathbb{Z}} \subseteq J.$$

where we assume without loss of generality that uncertainty only affects the coefficients a_{ij}. We model the uncertainty through a *multiband uncertainty set* $\mathcal{S}_M$, a natural generalization of Γ-Rob (see [6, 7] for a comparison between the two models). Specifically, we assume that for each coefficient a_{ij} we are given its nominal value $\bar{a}_{ij}$ and maximum negative and positive deviations $d_{ij}^{K^-}, d_{ij}^{K^+}$ from $\bar{a}_{ij}$. The actual value a_{ij} then lies in the interval $[\bar{a}_{ij} + d_{ij}^{K^-}, \bar{a}_{ij} + d_{ij}^{K^+}]$. We derive the generalization of Γ-Rob by partitioning the single deviation band $[d_{ij}^{K^-}, d_{ij}^{K^+}]$ for each coefficient a_{ij} into K bands, defined on the basis of K deviation values: $-\infty < d_{ij}^{K^-} < \cdots < d_{ij}^{-1} < d_{ij}^{0} = 0 < d_{ij}^{1} < \cdots < d_{ij}^{K^+} < +\infty$. We use these deviation values to define: (1) a set of positive deviation bands, such that each band $k \in \{1, \ldots, K^+\}$ corresponds to the range $(d_{ij}^{k-1}, d_{ij}^{k}]$; (2) a set of negative deviation bands, such that each band $k \in \{K^- + 1, \ldots, -1, 0\}$ corresponds to the range $(d_{ij}^{k-1}, d_{ij}^{k}]$ and band $k = K^-$ corresponds to the single value $d_{ij}^{K^-}$ (note that the interval of each band except $k = K^-$ is therefore open on the left). With a slight abuse of notation, we denote a generic deviation band by the index k, with $k \in K = \{K^-, \ldots, -1, 0, 1, \ldots, K^+\}$ and the corresponding range by $(d_{ij}^{k-1}, d_{ij}^{k}]$.

In order to complete the description of the uncertainty set, for each band $k \in K$, we introduce two values $l_k, u_k \in \mathbb{Z}^+$: $0 \leq l_k \leq u_k \leq n$, which respectively represent a lower bound and an upper bound on the number of deviations that may fall in k.

As additional assumptions, we do not limit the number of coefficients that may take their nominal value, i.e. $u_0 = n$, and we impose that $\sum_{k \in K} l_k \leq n$, to ensure the existence of a feasible realization of the coefficient matrix.

The robust counterpart of the program MILP can be defined by inserting in each constraint $i \in I$ the term $DEV_i(x, \mathcal{S}_M)$ that represents the maximum deviation allowed by the multiband uncertainty set for a solution x, (i.e., a robust constraint looks like $\sum_{j \in J} a_{ij}\, x_j + DEV_i(x, \mathcal{S}_M) \leq b_i$). The term $DEV_i(x, \mathcal{S}_M)$ is equal to the optimal value of a 0–1 linear maximization program that finds the worst coefficient distribution over the deviation bands for x (see [6] for details). The resulting robust counterpart is thus non-linear. However, using duality theory, we proved that this problem can be reformulated as a compact and linear problem, as stated in the following theorem (we refer the reader to [6, 7] for the formal complete statements and proofs of the theorems presented in this section).

Theorem 1 [Büsing and D'Andreagiovanni, 2012] *The robust counterpart of MILP under the multiband uncertainty set is equivalent to a compact mixed-integer linear program, which includes $K \cdot m + n \cdot m$ additional variables and $K \cdot n \cdot m$ additional constraints.*

In the case of large uncertain programs, the increase in size of the robust counterpart may represent an issue for obtaining a robust optimal solution quickly. We have thus investigated the possibility of developing a cutting-plane algorithm based on the separation of *robustness cuts*, that is, cuts that impose robustness. The basic question is simple: we are given a solution to the considered problem and we desire to check whether the solution is robust and feasible. If this is not the case, we separate a robustness cut and we add it to the problem, solving the new resulting problem. This step can be iterated as in a typical cutting-plane approach, until no robustness cut is needed and thus the generated solution is robust and optimal. In the case of Γ-Rob, the separation of a robustness cut is trivial and just consists in sorting the deviations and choosing the worst $\Gamma > 0$ [11]. This straightforward approach is not valid for multiband uncertainty, but we proved anyway that the separation can be done efficiently (see [6, 7] for the formal statement and the detailed description of how the min-cost flow instance is built):

Theorem 2 [Büsing and D'Andreagiovanni, 2012] *Under multiband uncertainty, the separation of a robustness cut for a constraint of MILP can be done in polynomial time by solving a* min-cost flow problem.

3 Multiband Robustness for 0–1 Programs

We now focus attention on the following 0-1 linear program:

$$\max \sum_{j \in J} c_j\, x_j \qquad \text{(BP)}$$

$$x_j \in X \subseteq \{0, 1\}^n \qquad j \in J,$$

in which the cost coefficients c_j are supposed to be non-negative (important optimization problems, like the shortest path problem and the minimum spanning tree problem, present this structure). Furthermore, we assume that the cost coefficients are subject to uncertainty and that uncertainty is modeled by a multiband set as follows: for each cost coefficient, we are given the nominal cost $\bar{c}_j$ and a sequence of $K^+ + 1$ deviation values d_j^k, with $k \in K = \{0, \dots, K^+\}$, such that $0 = d_j^0 < d_j^1 < \cdots < d_j^{K+} < \infty$ (we remark that in contrast to the previous section, we consider here without loss of generality only positive deviations). Through these values, we define: (1) the zero-deviation band corresponding to the single value $d_j^0 = 0$; (2) a set K^+ of positive deviation bands, such that each band $k \in K \backslash \{0\}$ corresponds to the range $(d_j^{k-1}, d_j^k]$. Finally, we introduce values $l_k, u_k \in \mathbb{Z}$, with $0 \leq l_k \leq u_k \leq n$, to represent the lower and upper bounds on the number of deviations falling in each band $k \in K$.

As BP is a special case of MILP, by Theorem 1, the compact and linear robust counterpart of BP is (see [7] for details):

$$\max \sum_{j \in J} c_j x_j + \sum_{k \in K} \theta_k w_k + \sum_{j \in J} z_j \qquad \text{(Rob-BP)}$$

$$w_k + z_j \geq d_j^k x_j \qquad j \in J, k \in K \qquad (1)$$

$$w_k, z_j \geq 0 \qquad j \in J, k \in K \qquad (2)$$

$$x_j \in X \subseteq \{0,1\}^n \qquad j \in J,$$

in which we note (i) the presence of additional non-negative variables (1); (ii) the presence of additional constraints (1); (iii) the insertion of additional terms in the objective function. The coefficients $\theta_k \geq 0$ constitute the so-called *profile* of the multiband uncertainty set and are equal to the number of coefficients that must fall in the band k to maximize the deviation (the values θ_k are derived from the values l_k, u_k exploiting domination between feasible realizations of the uncertainty set [7]).

A robust optimal solution can be computed by solving *Rob-BP* or by adopting the cutting-plane approach based on robustness cuts and presented in the previous section. Anyway, as an alternative to these two general approaches, we proved the following special result (see [7] for details):

Theorem 3 *The robust optimal solution of BP with cost uncertainty modeled through a multiband set can be computed by solving a polynomial number of nominal problems BP with modified objective function, if the number of bands is constant. Tractability and approximability of BP are maintained.*

In addition to these results, we characterized a new family of valid inequalities for the robust counterpart of BP, by adopting a proof strategy similar to that of Atamtürk for Γ-Rob [1] (see [10] for the proof).

Proposition 1 *For any $k \in K$ and subset $T = \{j_1, j_2, \dots, j_t\} \subseteq J$ with $0 = d_{j_0}^k \leq d_{j_1}^k \leq \cdots \leq d_{j_t}^k$, the following inequality is valid for problem (Rob-BP):*

$$\sum_{j_l \in T} \left(d^k_{j_l} - d^k_{j_{l-1}} \right) x_{j_l} \leq w_k + \sum_{j \in T} z_j$$

Additionally, if $0 = d_{j_0} < \cdots < d_{j_t}$*, then the previous inequalities are facet-defining.*

4 Robust Wireless Network Design

We used our new results about uncertain 0–1 programs in a set of preliminary experiments considering a central problem in wireless network design: the *power assignment problem* (PAP). The PAP considers the design of a wireless network made up of a set of transmitters T providing a telecommunication service to a set of users U. It essentially consists of setting power emissions of the transmitters, while minimizing a function of emitted powers. For an exhaustive introduction to the wireless network design problem and to the PAP, we refer the reader to [8, 12]. The PAP has recently regained attention, due to ongoing switches from analog to digital television that have taken place in many countries over the last years. Here, we consider a variant of the PAP that has been recently brought to our attention from our partners in former industrial cooperations: instead of minimizing the simple summation of the power emission, we multiply the power of each transmitter by the price paid to buy the power (big network operator can indeed profit from special energy fees, which usually vary from transmitter to transmitter). This variant of the PAP can be modeled as follows: we use a vector of non-negative continuous variables p to represent the power emissions of transmitters. Then we introduce (1) a vector π to represent the price of a energy unit for each transmitter, (2) a matrix A to represent signal attenuation for each transmitter-user pair, (3) a vector δ to represent the minimum power that guarantees service coverage for a user (signal-to-interference threshold). Using these elements, the PAP can be written in the following matrix form: $\min_p \{ \pi' p : Ap \geq \delta,\ p \geq 0^{|U|} \}$. Because of the presence of the attenuation coefficients that may vary in a very wide range, this formulation is known to lead to numerical instability in the solution process, which may greatly reduce the effectiveness of commercial optimization solvers. As a remedy, in our computational study, we have considered a tighter pure 0–1 formulation, the so-called *power-indexed formulation*, based on the use of discrete power variables and of a special family of *generalized upper bound cover inequalities* (see [8, 9] for details).

The energy price coefficients of the objective function are supposed to be subject to uncertainty: a big wireless operator can indeed establish energy contracts based on favorable prices that may however fluctuate (within limits) due to load conditions of the energy network and to variability of price formation dynamics of the energy market. Under these conditions, professionals would be interested in getting robust solutions to the PAP, namely power configurations satisfying the coverage constraints, while minimizing the total power expense and taking into account price

deviations specified by an uncertainty set reflecting their risk aversion. If the uncertainty is modeled by a multiband set, the resulting robust counterpart of the problem can be solved by adopting the sequential solution approach sketched in Theorem 3: indeed, we face a (pure binary) power-indexed formulation of the PAP, where the uncertainty only affects the price coefficients in the objective function. Based on a series of discussions with experts, we suppose that each price coefficient is distributed according to a histogram resembling the shape of an exponential distribution. The adoption of the Bertsimas-Sim Γ-robustness model would provide a low-resolution modeling of such histogram. The multiband uncertainty model grants in contrast a more accurate representation that reduces conservatism. Based on discussions aimed at pointing out the risk aversion of the professionals, we adopted a system of 5 deviation bands. Our experiments considered a set of 15 realistic network instances of increasing size (including up to 150 users and 10 transmitters), all based on the WiMAX technology. All the experiments were made on a 2.70 GHz machine with 8 GB RAM and using IBM ILOG Cplex 12.1 as optimization solver.

The main purpose of our tests was to compare the efficiency of solving directly the compact formulation (Rob-BP) with that of the sequential approach sketched in Theorem 3 and formalized in [7]. The sequential approach performed slower in the case of 5 instances, while in all the other cases it reduced the solution time of 12 % on average. Taking into account the computational challenge of a power-indexed PAP, we consider this reduction significative and we believe that it could be enhanced by a smart exploitation of the new family of (strong) valid inequalities identified in Proposition 1, which will be the object of future experimentation. From the point of view of the price of robustness, the refined representation of the uncertainty granted by the multiband set guaranteed a reduction of up to 15 % in the conservatism of the robust optimal solution with respect to Γ-robustness (thus sensibly reducing the increase in power expense that a network operator must face to protect against price fluctuations).

References

1. Atamtürk, A. (2006). Strong formulations of robust mixed 0–1 programming. *Mathematical Programming B*, *108*, 235–250.
2. Bauschert, T., Büsing, C., D'Andreagiovanni, F., Koster, A. M. C. A., Kutschka, M., Steglich, U. (2014). Network planning under demand uncertainty with robust optimization. To appear in *IEEE Communications Magazine*, *52*(2), 178–185.
3. Ben-Tal, A., El Ghaoui, L., Nemirovski, A. (2009). Robust optimization. Berlin: Springer.
4. Bertsimas, D., Brown, D., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, *53*(3), 464–501.
5. Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research*, *52*(1), 35–53.
6. Büsing, C., D'Andreagiovanni, F. (2012). New results about multi-band uncertainty in robust optimization. In: R. Klasing (Ed.), Experimental algorithms - SEA 2012. Berlin: Springer.
7. Büsing, C., D'Andreagiovanni, F. (2012). Robust optimization under multiband uncertainty—Part I: Theory (preprint: Optimization Online 13–01-3748).
8. D'Andreagiovanni, F. (2012). Pure 0–1 programming approaches to wireless network design. *4OR: A Quarterly Journal of Operations Research*. doi:10.1007/s10288-011-0162-z.

9. D'Andreagiovanni, F., Mannino, C., & Sassano, A. (2013). GUB covers and power-indexed formulations for wireless network design. *Management Science*, *59*(1), 142–156.
10. D'Andreagiovanni, F., Raymond, A. (2013). Multiband robust optimization and its adoption in harvest scheduling. Forthcoming in Proceedings of FORMATH 2013, Fukushima, Japan.
11. Fischetti, M., & Monaci, M. (2012). Cutting plane versus compact formulations for uncertain (integer) linear programs. *Mathematical Programming Computation*, *4*(3), 239–273.
12. Mannino, C., Rossi, F., & Smriglio, S. (2006). The network packing problem in terrestrial broadcasting. *Operations Research*, *54*(6), 611–626.

A Branch-and-Price Approach for a Ship Routing Problem with Multiple Products and Inventory Constraints

Rutger de Mare, Remy Spliet and Dennis Huisman

Abstract In the oil industry, different oil products are blended in a refinery. Afterwards, these products are transported to different harbors by ship. Due to the limited storage capacity at the harbors and the undesirability of a stock-out, inventory levels at the harbors have to be taken into account during the construction of the ship's routes. In this paper, we give a detailed description of this problem, which we call the ship routing problem with multiple products and inventory constraints. Furthermore, we formulate this problem as a generalized set-covering problem. We propose a branch-and-price algorithm to solve it and we discuss this briefly.

1 Introduction

At a refinery crude oil is separated into different components. These components are blended into oil products called grades, which are stored in tanks. From the product tanks, grades are transported via primary terminals and secondary terminals to the customer. The transport from the refinery to the primary terminals located at a harbor is often done by ship. The large scale of the transported quantities implies that a lot of money is involved. The operating cost of a ship are typically between \$50,000 and \$400,000 a day. The inventory cost (including working capital cost) of

R. de Mare (✉)
ORTEC, P.O. Box 75, 2719 EA Zoetermeer, The Netherlands
e-mail: rutger.demare@ortec.com

R. Spliet · D. Huisman
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738,
3000 DR Rotterdam, The Netherlands
e-mail: spliet@ese.eur.nl

D. Huisman
e-mail: huisman@ese.eur.nl

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_14,

one grade in one harbor is typically around $1 million per year. A trade off should be made between inventory costs, operating costs of different types of ships and the risk of a stock-out. Ideally, ship routing decisions and inventory decisions should be a combined decision.

In this paper, we consider a ship routing problem with multiple products and inventory constraints. Ronen [6] presents a two-stage heuristic for the multiple product ship-scheduling problem. Al-Khayyal and Hwang [1] formulate the problem as a mixed integer linear program. They show that even small instances cannot be solved by a general purpose solver and they argue that specialized algorithms are needed. A specialized exact algorithm is presented by Christiansen [3] in the case of a single-product ship scheduling problem, and later by Brønmo et al. [2] for a ship scheduling problem with flexible cargo sizes. Christiansen [3] looks at the transportation of ammonia with production and consumption factories. In our case of oil transportation, multiple products are considered, namely several grades of oil. The different grades are transported in different compartments of a ship, and stored in different tanks at the harbor.

In Sect. 2, we give a description of the ship routing problem. We formulate the problem by extending the formulation of [4]. In Sect. 3, we discuss a Branch-and-Price method to solve the problem. We finish the paper with concluding remarks in Sect. 4.

2 Problem Description

Let G be the set of grades. Furthermore, let H be the set of harbors. Demand at harbor $h \in H$ for grade $g \in G$ is linear in time with rate d_{hg}. Note that for a harbor h where grade g is produced $d_{hg} < 0$. The storage capacity at harbor $h \in H$ for grade $g \in G$ is given by Q_{hg}. Note that the inventory of a grade may never exceed the capacity. A set of vessels V is available to transport grades between harbors. They are used to ensure that the inventory constraints at each harbor are never violated during a finite planning horizon of length T. Each vessel $v \in V$ consists of a set of compartments C_v that are used to store the grades. We denote the first compartment $c \in C_v$ by c_v^1. We assume that each compartment may contain every type of grade and no cleaning is required when first transporting one type of grade in a compartment and then another. Let Q_c be the capacity of compartment c. For vessel $v \in V$, travel costs between harbors $i \in H$ and $j \in H$ are given by c_{ij} and travel time is given by t_{ij}. Furthermore, the rate of unloading/loading of grade $g \in G$ at harbor $h \in H$ by vessel $v \in V$ is r_{vhg}. For each harbor $h \in H$ an ordered set of non-overlapping time windows $W_h = \{w_1, \ldots, w_k\}$ is available. We denote each time window $w \in W$ as $[\underline{w}, \overline{w}]$, where $\underline{w}$ is the start of the time window and $\overline{w}$ is the end of the time window. We allow at most one vessel to visit harbor h during one time window. Also, we allow the vessel to wait.

The ship routing problem with multiple products and inventory constraints, SRPMPIC, is the problem of scheduling the vessels to transport grades between

harbors to ensure that the inventory constraints are not violated during the planning horizon, such that the total transportation costs are minimized. Before presenting a mixed integer linear programming formulation for the SRPMPIC, we discuss two underlying networks of this problem, the harbor-grade network and the compartment network. These networks are used in our formulation.

2.1 The Harbor-Grade Network

The harbor-grade network for harbor $h \in H$ and grade $g \in G$ is used to represent a loading or unloading schedule of grade g at harbor h. For ease of exposition, we will consider loading as unloading a negative amount. During each time window $w \in W_h$ an amount of grade g is possibly unloaded at harbor h.

For harbor h, we construct a set of nodes N consisting of pairs (w, q) for $w \in W_h$ and q representing the cumulative amount of grade g delivered up to and including time window w. To limit the size of this set of nodes, we consider only a finite number of unloaded quantities during each time window, including the amount 0. For example, a vessel is allowed to unload exactly 0, 20 or 40 units of grade g during a single time window. Furthermore we construct a set of arcs A connecting nodes $(w, q) \in N$ and $(w', q') \in N$ if w is the direct predecessor of w' in W_h and $q' - q$ is exactly one of the allowed unloading amounts.

The harbor-grade network is denoted as the acyclic graph $N_{hg} = (N, A)$. An unloading schedule is defined as a pair (P, t), where P is a path in N_{hg} from $(w_1, 0)$ to (w_k, q), for w_1 the first and w_k the last time window in W_h and q such that $(w_k, q) \in N$, and t is a vector containing the time of service at each location on the path.

An unloading schedule is feasible if (i) for every (w, q) on P and corresponding arrival time t unloading of the entire quantity can be completed between t and $\overline{w}$ and (ii) the inventory constraints are not violated at any time during the planning horizon. Note that the inventory constraints might be modelled by adjusting the time window w corresponding to node $(w, q) \in N$ to disallow unloading too early or too late resulting in a violation of the inventory constraints. We denote the set of all feasible harbor-grade schedules at harbor $h \in H$ and grade $g \in G$ as S_{hg}.

2.2 The Compartment Network

The compartment network for compartment $c \in C_v$ of vessel $v \in V$ is used to represent a sailing schedule of compartment c to load and unload different grades at different harbors.

We construct a set of nodes $\hat{N}$ consisting of tuples $(h, w, q_1, \ldots, q_{|G|})$ for every $h \in H$, for every $w \in W_h$ and for $q_1 \ldots, q_{|G|}$ representing the total load of the grades in G. As in the harbor-grade network, we consider only a finite number of unloading

and loading quantities including the amount 0. Also, note that we only consider tuples in which the load of at most one grade is nonzero, as a compartment may only carry one type of grade at any moment in time. Furthermore, we construct a set of arcs $\hat{A}$ connecting nodes $(h, w, q_1, \ldots, q_{|G|}) \in \hat{N}$ and $(h', w', q'_1, \ldots, q'_{|G|}) \in \hat{N}$ if harbor h' can be reached during time window w' when departing from harbor h during time window w, and if the $q'_g - q_g$ is exactly one of the allowed unloading amounts for all $g \in G$. To each arc we associate travel time of sailing from one harbor to the other. To each arc we also assign travel costs (we elaborate on these costs later).

The compartment network is denoted as the cyclic graph $\hat{N}_c = (\hat{N}, \hat{A})$. A compartment schedule is defined as a pair (P, t) where P is a path in $\hat{N}_c$ and t is a vector containing the time of service at each location on the path. A compartment schedule is feasible if the time windows at each node on the path are not violated. We denote the set of all feasible compartment schedules for compartment c as R_c.

2.3 Problem Formulation

Next, we formulate the SRPMPIC as a set partitioning problem. It is an extension of the formulation by [4] for the single product ship scheduling problem.

With each compartment schedule $r \in R_c$ we associate the variable x_r indicating whether compartment schedule r is used. Let the costs of schedule r be given by p_r. For the routes $R_{c_v^1}$ of the first compartment of every vessel c_v^1, these costs are the actual traveling costs. For all other compartments these costs are 0. In our formulation we link the routes of the different compartments of one vessel. Therefore, the traveling costs of a vessel are represented in this way. Furthermore, we associate with each compartment schedule $r \in R_c$ the parameter q_{rhgw} representing the amount of grade g unloaded at harbor h during time window w. Also, we associate with r the parameter t_{rhgw} specifying the time at which a ship commences unloading of grade g at harbor h during time window w. Let $t_{rhgw} = 0$ if no unloading is performed.

We allow x_r to be any fractional value between 0 and 1, hence convex combinations of compartment schedules might be selected. However, we impose integrality on the physical route that is traveled. This way, any quantity of grade g may be unloaded at harbor h during time window w, in particular not only the allowed quantities used in the construction of R_c. To model this, we introduce an auxiliary variable $z_{hwh'w'c}$ that indicates whether compartment c departs from harbor h during time window w and travels to harbor h' to arrive during time window w'. Furthermore, we introduce the parameter $a_{hwh'w'r}$ indicating whether route r departs from harbor h during time window w and travels to harbor h' to arrive during time window w'.

With each harbor-grade schedule $s \in S_{hg}$ we associate the variable y_s indicating whether schedule s is used. We associate with each harbor-grade schedule $s \in S_{hg}$ the parameter q_{sw} representing the amount of grade g unloaded at harbor h during time window w. Furthermore, we associate with s the parameter t_{sw} specifying the

time at which unloading of grade g commences at harbor h during time window w. Let $t_{sw} = 0$ if no unloading is performed.

Similar to x_r we allow y_s to be any fractional value between 0 and 1. However, there is no need to impose integrality conditions, as any convex combinations of harbor-grade schedules is a feasible harbor-grade schedule in which any quantity may be unloaded during every time window, while still satisfying the inventory constraints. The SRPMPIC can be formulated as follows:

$$\min \sum_{\substack{v \in V \\ r \in R_{c_v^1}}} p_r x_r \tag{1}$$

$$\sum_{\substack{v \in V \\ c \in C_v \\ r \in R_c}} q_{rhgw} x_r - \sum_{s \in S_{hg}} q_{sw} y_{sw} = 0 \quad \forall h \in H, g \in G, w \in W_h \tag{2}$$

$$\sum_{\substack{v \in V \\ r \in R_{c_v^1}}} t_{rhgw} x_r - \sum_{s \in S_{hg}} t_{sw} y_{sw} = 0 \quad \forall h \in H, g \in G, w \in W_h \tag{3}$$

$$\sum_{r \in R_{c_v^1}} t_{rhgw} x_r - \sum_{r \in R_c} t_{rhgw} x_{rhgw} = 0 \quad \forall v \in V, h \in H, g \in G, w \in W_h, c \in C_v \backslash c_v^1 \tag{4}$$

$$\sum_{r \in R_c} x_r = 1 \quad \forall v \in V, c \in C_v \tag{5}$$

$$\sum_{s \in S_{hg}} y_s = 1 \quad \forall h \in H, g \in G \tag{6}$$

$$\sum_{r \in R_c} a_{hwh'w'r} x_r - z_{hwh'w'c} = 0 \quad \forall h, h' \in H, w, w' \in W_h, v \in V, c \in C_v \tag{7}$$

$$x_r \in [0, 1] \quad \forall r \in R_c, c \in C_v, v \in V \tag{8}$$

$$y_s \in [0, 1] \quad \forall s \in S_{hg}, h \in H, g \in G \tag{9}$$

$$z_{hwh'w'c} \in \{0, 1\} \quad \forall h, h' \in H, w, w' \in W_h, v \in V, c \in C_v \tag{10}$$

The objective function is represented by (1). Constraints (2) ensure that the total quantity of grade g unloaded by the vessels at harbor h during time window w is equal to the quantity as specified in the selected harbor-grade schedule. Similarly, constraints (3) ensure that the unloading of grade g by the vessels commences at harbor h during time window w as specified in the selected harbor-grade schedule. Furthermore, constraints (4) ensures that all compartments of a single vessel travel the same route. Constraints (5) and (6) ensure that for each compartment and for

every harbor-grade combination, exactly one schedule is selected. Finally, integrality constraints are given by (7) and (10).

3 Branch-and-Price

We suggest solving the SRPMPIC using a branch-and-price algorithm. Lower bounds can be found by solving the LP relaxation of formulation (1)–(10). As the formulation contains many variables, we suggest using a column generation algorithm to solve the LP relaxation. Initially we consider a restricted master problem, which is the SRPMPIC including only a limited number of compartment schedules and harbor-grade schedules. We iteratively solve the restricted master problem and add new compartment schedules and harbor-grade schedules by solving a pricing problem. The restricted master problem is solved using the simplex algorithm, yielding dual multipliers corresponding to each constraint. Next the pricing problems are solved to identify schedules with negative reduced costs. If such schedules are found, they are added to the restricted master problem. If none exist, the solution to the current restricted master problem is also the optimal solution to the LP relaxation of (1)–(10).

The pricing problem decouples per harbor-grade combination and compartment. The pricing problem for generating a harbor-grade variable for harbor h and grade g is a shortest path problem in N_{hg} with time window constraints and linear node costs. It can be solved using the labeling algorithm by [5]. The pricing problem for generating a compartment variable for compartment c is an elementary shortest path problem in the cyclic graph $\hat{N}_c$ with time window constraints and linear node costs. Note that in particular cyclic paths in which a harbor is visited during the same time window more than once is undesired. Nonetheless, we relax the elementarity condition, and allow cyclic schedules to be generated. This lowers the value of the LP bound while the optimal integer solution remains the same, as such schedules cannot be part of any integer solution. With this relaxation, the pricing problem is a shortest path problem with time window constraints and linear node costs. It can be solved using the same labeling algorithm by [5] as used to solve the harbor-grade pricing problem.

4 Concluding Remarks

In this paper, we discussed the ship routing problem with multiple products and inventory constraints which arises in the oil industry. Since different oil products cannot be stored in the same compartment of a ship, the capacities of the different compartments of the ship have to be taken into account as well. We have extended the formulation and branch-and-price algorithm by [4] to the case of multiple products and multiple compartments.

References

1. Al-Khayyal, F., & Hwang, S. J. (2007). Inventory constrained maritime routing and scheduling for multi-commodity liquid bulk, part I: Applications and model. *European Journal of Operations Research*, *176*, 106–130.
2. Brønmo, G., Nygreen, B., & Lysgaard, J. (2009). Column generation approaches to ship scheduling with flexible cargo sizes. *European Journal of Operations Research*, *200*, 139–150.
3. Christiansen, M. (1999). Decomposition of a combined inventory and time constraint ship routing problem. *Transportation Science*, *33*, 3–16.
4. Christiansen, M., & Nygreen, B. (2005). Robust inventory ship routing by column generation. In G. Desaulniers, J. Desrosiers, & M. M. Solomon (Eds.), *Column Generation* (pp. 197–224). New York: Springer.
5. Ioachim, I., Gélinas, S., Soumis, F., & Desrosiers, J. (1998). A dynamic programming algorithm for the shortest path problem with time windows and linear node costs. *Networks*, *31*, 193204.
6. Ronen, D. (2002). Marine inventory routing: shipments planning. *Journal of the Operations Research Society*, *53*, 108–114.

Data Driven Ambulance Optimization Considering Dynamic and Economic Aspects

Dirk Degel, Lara Wiesche and Brigitte Werners

Abstract Providing high quality emergency medical services (EMS) and ensuring accessibility to these services for the general public is a key task for health care systems. Given a limited budget available resources, e.g. ambulances, have to be used economically in order to ensure a high quality coverage. Emergency vehicles have to be positioned and repositioned such that emergencies can be reached within a legal time frame. Empirical studies have shown temporal and spatial variations of emergency demand as well as variations of travel times during a day. The numbers of emergency calls within 24 h differ significantly between night and day and show peaks especially during rush hours. We provide a data driven model considering time and spatial dependent degrees of coverage. This allows a simultaneous optimization of empirically required coverage with minimal number of ambulances, respectively costs. Therefore utilization and quality criteria are to be implemented. An integer linear program is formulated using time periods in order to model time-dependent demand and time-dependent travel times. It is shown on large empirical data records that the presented dynamic model outperforms existing static models with respect to coverage and utilization of resources.

D. Degel (✉) · L. Wiesche · B. Werners
Faculty of Management and Economics, Chair of Operations Research and Accounting, Ruhr University Bochum, 44780 Bochum, Germany
e-mail: dirk.degel@rub.de

L. Wiesche
e-mail: lara.wiesche@rub.de

B. Werners
e-mail: or@rub.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_15,

1 Introduction

Providing high quality medical services and ensuring accessibility to these services for the general public is a key task for a health care system. Given a limited budget available resources, e. g. ambulances or locations of EMS and fire departments, have to be planned and used economically in order to ensure high quality supply [2–4]. Explicitly, during a regular day EMS-vehicles have to be positioned and re-positioned such that emergencies can be reached within a legal time frame. Empirical studies show that demand changes over time and that there are regional differences. In the current situation in Bochum ambulances are placed at existing EMS and fire departments. Because these rescue departments are located near the city center, this leads to a very high degree of coverage in the city center and causes undersupply in peripheral areas. In particular some demand areas are covered ninefold and in many cases far exceeds the required degree of coverage. In contrast peripheral regions are covered only once and some of these regions are not covered at all within a given time limit. In order to handle these effects, the required or necessary coverage is investigated empirically. An integer linear program (ILP) is applied in order to locate and relocate ambulances according to a required degree of coverage. For this, a number of additional, flexible ambulance locations will be considered. The goal is to use resources such as ambulances efficiently and ensure the empirically determined necessary coverage. This leads to a high level of service and at the same time avoids over-coverage and saves resources.

2 Identifying Empirically Necessary Coverage

A large number of models have been developed in order to support decision making for ambulance location in various decision situations. Farahani et al. [5] provide a comprehensive survey of covering models which are typical for EMS applications and Li et al. [8] provide a well structured survey of optimization models with focus on emergency response applications. Additionally, Başar et al. [1] and Hulshof et al. [7] give taxonomic overviews of decision support systems. In almost all presented models a unique degree of coverage is maximized. For example, a double-coverage is considered in the *Double Standard Model* (DSM) by Gendreau et al. [6] and its extensions [9]. Instead of ensuring double coverage for each demand region during the entire day analytics and data driven optimization can be used to determine a better level of necessary coverage. We investigate empirically the number of emergency situations occurring simultaneously in order to determine the required degree of coverage (see Fig. 1). Each demand site is analyzed individually due to the fact that usually observed demand is not equally distributed over the planning area. To calculate the necessary coverage degree $\mathrm{e}(i)$ of a demand node i we have to ensure that the probability that an emergency call could not be served because no ambulance is available is less then $1 - \beta\,\% = 5\,\%$, or in other words, that:

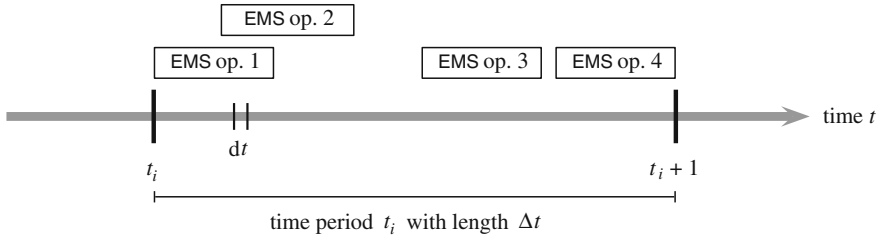


Fig. 1 Consideration of parallel emergency operations. In this situations a double coverage (two ambulances) is necessary

$$\textit{Probability} \left\{ \begin{array}{c} \text{\# of ambulances can cover} \\ \text{demand node } i \end{array} \geq \begin{array}{c} \text{\# of parallel emergencies in} \\ \text{the area around } i \end{array} \right\} \geq 95\,\%$$

First a static version of the model which maximizes the empirically determined necessary coverage is formulated and then we consider dynamic, time-dependent modifications.

2.1 Model with Empirically Necessary Coverage

The (standard) DSM seeks to maximize the demand covered twice within a time standard of r_1, using p ambulances and subject to the double covering constraints. In our approach the static model maximizes the demand, which is covered $\mathrm{e}(i)$-times:

$$\max \quad \sum_{i \in I} d_i x_i^{\mathrm{e}(i)} \tag{1}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{N}_i^{r2}} y_j \geq 1 \qquad \forall i \in I \tag{2}$$

$$\sum_{i \in I} d_i x_i^1 \geq \alpha \sum_{i \in I} d_i \qquad \forall i \in I \tag{3}$$

$$x_i^{k-1} \geq x_i^k \qquad \forall i \in I,\ \forall k \in \{2, \ldots, p\} \tag{4}$$

$$\sum_{j \in \mathcal{N}_i^{r1}} y_j \geq \sum_{k=1}^{p} x_i^k \qquad \forall i \in I \tag{5}$$

$$\sum_{j \in J} y_j \leq p \tag{6}$$

$$x_i^k \in \{0, 1\} \qquad \forall i \in I,\ \forall k \in \{1, \ldots, p\} \tag{7}$$

$$y_j \in \mathbb{N}_0 \qquad \forall j \in J \tag{8}$$

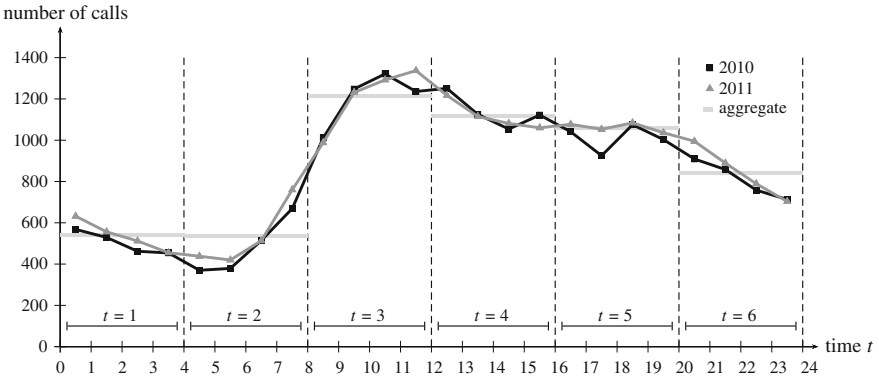


Fig. 2 Number of emergency calls for a 24-h-day in a German mid-size city and average (aggregated) speed

where d_i is the demand at node $i \in I$, $\mathcal{N}_i^{r_\ell} := \{j \in J \mid t_{ij} \leq r_\ell\}$ for $r_1 < r_2$ characterises the neighborhood sets of demand node i and p represents the total number of ambulances. The decision variable

$$x_i^k := \begin{cases} 1, \text{ if demand node } i \text{ is covered } k \in \{1, \ldots, p\} \text{ times} \\ 0, \text{ else.} \end{cases}$$

y_j represents the number of ambulances located at node j. The objective function computes the demand covered e(i)-times within r_1 time units. The combination of constraints (2) and (3) ensures that a proportion α of the total demand is covered within r_1 and the whole demand area is covered within r_2. Constraints (3) and (4) express the necessary coverage requirements. The left-hand side of (5) represents the number of ambulances covering demand node i within r_1 time-units, while the right-hand side is 1 if i is covered once and so on within r_1. Equation (6) limits the number of ambulances to p. (7) and (8) describe the domain of the decision variables.

2.2 Time-Dependent Considerations

In addition to considering empirical necessary coverage e(i), significant time-dependent variations in the input parameters as demand, travel-time and necessary coverage can be observed. Almost all models in literature do not include all dynamic aspects at the same time. In Fig. 2 the number of emergency calls is indicated with respect to every hour of the day. The figure shows that there are significant differences in demand between night, day and peaks especially during rush hours. Moreover, this figure clearly indicates that demand changes during the day. A required constant degree of coverage will either underestimate or overestimate actual demand, e. g. e(i, t) is time-dependent. However, existing models do not

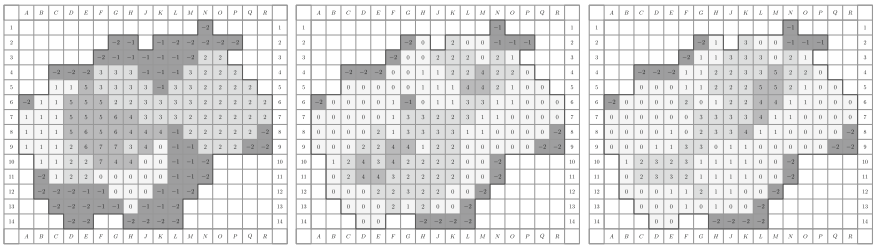


Fig. 3 Differences between the required empirical coverage and the coverage obtain from (*1*) status quo, (*2*) solution by the double covering maximization model, and (*3*) our solution by the empirically determined suitable covering maximization model for rush hour period (8–12 a.m.)

consider time-dependency of model parameters such as demand and travel times for ambulances. A new modeling approach is developed that explicitly integrates demand and travel times varying simultaneously throughout the day. In order to generate more flexibility in the EMS-system, we allow the assignment of ambulances not only to existing EMS-departments but also to additional, flexible locations such as hospitals or volunteer fire departments. Variations in the fleet size during the day depending on changes in travel speed are explicitly included to consider economic aspects. Dynamic allocation of ambulances at additional, flexible locations and relocations to the main EMS-departments are required to handle time-dependent changes in travel-speed and demand. The degree of coverage, the number of relocations and the fleet size are considered to be major performance indicators. Incorporating these aspects leads to a dynamic version of model (1)–(8).

3 Improvement of Status Quo

The dynamic model is part of a decision support tool that is developed for urban emergency services. The aim is to support strategic and tactical decisions. The following figures clearly show the improvement of the status quo. For Bochum (Germany) it can be seen, that the new model also outperforms the double coverage model. Figure 3 illustrates the positive effects of maximizing the empirical coverage for a time period around midday in which demand is typically high (see Fig. 1). The three maps depict the deviation from empirical coverage to (1) the actual solution ("status quo") which is applied by the EMS in Bochum, Germany, (2) a solution determined by dynamic model similar to the model presented by Schmid and Doerner [9] with a double coverage optimization function and the solution of our new model (3). The evaluation considers the deviation of necessary coverage and the coverage obtain by the status quo or the models. White squares mean that the empirically necessary degree of coverage is achieved by the solution. Attaining the empirical level and also small positive differences are preferable. Besides a very low level of coverage (dark gray) which can lead to non-sufficient supply of population, also a very high

degree of coverage (light gray) should not be tolerated because it wastes resources that could be utilized in a better way. The current situation shows typical results for urban areas: planning sites in the city center are "over-covered" to a large extent (more than 7-times over the necessary level). Yet, the resulting degree of coverage in the periphery is very often below a target level. The improvements according to integrating time-dependent and spatial demand become obvious. Data driven optimization and analytic methods as well as dynamic considerations lead to an efficient ambulance utilization. The same service level in the system can be ensured by less ambulances.

4 Conclusions

An evaluation using real-world data from 2010 to 2012 clearly points out that considering time-dependent travel times and time-dependent demand in our approach outperforms existing solutions using static model parameters. Overall, the proposed approach leads to a high quality solution with respect to coverage and cost criteria.

Acknowledgments This research is financially supported by *Stiftung Zukunft NRW*. The authors are grateful to staff members of the *Feuerwehr und Rettungsdienst Bochum* for detailed insights.

References

1. Başar, A., Çatay, B., & Ünlüyurt, T. (2011). A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. *Journal of the Operational Research Society*, *64*(4), 627–637.
2. Degel, D., Rachuba, S., Wiesche, L., & Werners, B. (2014). Reorganizing an existing volunteer fire station network in Germany. *Socio-Economic Planning Sciences* (in press). doi:10.1016/j.seps.2014.03.001. http://www.sciencedirect.com/science/article/pii/S0038012114000147.
3. Degel, D., Wiesche, L., Rachuba, S., Werners, B. (2013). Dynamic ambulance location providing suitable coverage for time-dependent demand. In: Gunal T., Gunes, M. M., Cayirli, E. D., Ormeci, E. L. (ed), *Operational Research Applied to Health Services (ORAHS) 2013 Conference Proceedings*.
4. Degel, D., Wiesche, L., Rachuba, S., & Werners, B. (2014). *Time-dependent ambulance allocation considering data driven empirical necessary coverage*. Health Care Management Science, 1–15 (in press). doi:10.1016/j.seps.2014.03.001. http://dx.doi.org/10.1007/s10729-014-9271-5.
5. Farahani, R., Asgari, N., Heidari, N., Hosseininia, M., & Goh, M. (2012). Covering problems in facility location: A review. *Computers & Industrial Engineering*, *62*(1), 368–407.
6. Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science*, *5*(2), 75–88.
7. Hulshof, P. J. H., Kortbeek, N., Boucherie, R. J., & Hans, E. W. (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems*, *1*(2), 129–175.

8. Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research*, *74*(3), 281–310.
9. Schmid, V., & Doerner, K. (2010). Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, *207*(3), 1293–1303.

Risk-Adjusted On-line Portfolio Selection

Robert Dochow, Esther Mohr and Günter Schmidt

Abstract The objective of on-line portfolio selection is to design provably good algorithms with respect to some on-line or offline benchmark. Existing algorithms do not consider 'trading risk'. We present a novel risk-adjusted portfolio selection algorithm (RAPS). RAPS incorporates the 'trading risk' in terms of the maximum possible loss. We show that RAPS performs *provably* 'as well as' the Universal Portfolio (UP) [4] in the worst-case. We *empirically* evaluate RAPS on historical NYSE data. Results show that RAPS is able to beat BCRP as well as several 'follow-the-winner' algorithms from the literature, including UP. We conclude that RAPS outperforms in case the assets in the portfolio follow a positive trend.

1 On-line Portfolio Selection

Let P denote the on-line portfolio selection algorithm, and let OPT denote the optimal offline algorithm. The input sequence becomes available to P over time, while OPT knows the whole input sequence in advance. The performance of P is evaluated by means of worst-case competitive analysis without making any statistical assumptions, e.g., on the nature of the stock market. The outcome is the ratio between the value obtained by OPT and P on a worst-case instance, called *regret*.

More formally, on-line portfolio selection aims to determine practical P for investing wealth among a set of m assets ($i = 1, \ldots, m$) over T trading periods ($t = 1, \ldots, T$). The finance community mainly addresses the problem of

R. Dochow (✉)
Saarland University, Saarbrücken, Germany
e-mail: rd@orbi.uni-saarland.de

E. Mohr
University of Mannheim, Mannheim, Germany
e-mail: mohr@bwl.uni-mannheim.de

G. Schmidt
Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_16,

maximizing the risk-adjusted return, while the information theory and machine learning community aims to maximize the terminal wealth $W_T(P)$ of P. The output of any P is a sequence of allocation vectors $\mathbf{b} = b_1, \ldots, b_T$ for the m assets, with $b_t = (b_{t1}, \ldots, b_{tm})$. The elements b_{ti} represent the proportion of wealth to be invested in the ith asset at the beginning of the tth period ($\sum_{i=1}^{m} b_{ti} = 1$). Let q_{ti} be the price of asset i at time t, and let $\mathbf{x} = x_1, \ldots, x_T$ denote an arbitrary sequence of m-dimensional price relative vectors x_t of the m assets over T. Then the elements of x_t are positive price relatives $x_{ti} = q_{ti}/q_{t-1i}$ of the ith asset at the end of the tth period. In other words, within the tth period the *portfolio return* in-/decreases by the factor $b_t^{\mathrm{T}} x_t = \sum_{i=1}^{m} b_{ti} x_{ti}$. Thus, after T trading periods, the terminal wealth achieved by P equals $W_T(P) = W_o \prod_{t=1}^{T} b_t^{\mathrm{T}} x_t$, where W_o denotes the initial wealth and is set to \$1 for convenience in this work.

In general, any P usually *learns* to compete with a target set of N reference algorithms ($j = 1, \ldots, N$). Let $\mathtt{Q} = \left\{Q^1, \ldots, Q^N\right\}$ denote this set. Following the concept of competitive analysis, the performance of P is measured by the *worst-case logarithmic wealth ratio* [3, p. 278]

$$\mathbf{W}_T(P, \mathtt{Q}) = \sup_{\mathbf{x}} \sup_{Q \in \mathtt{Q}} \ln \frac{W_T(Q)}{W_T(P)}, \tag{1}$$

where $\mathtt{Q}$ can be chosen arbitrarily. Most common is the class of constant-rebalanced portfolio (CRP) algorithms, or a mixture of different classes of algorithms.

CRP maintains a constant fraction of the total wealth in each of the underlying m assets. In an *i.i.d.* market if T is large, then *OPT* is the Best CRP (BCRP) [2]. Thus, on-line portfolio selection always chooses a target set $\mathtt{B} = \left\{B^1, \ldots, B^N\right\}$ of N CRP reference algorithms, known as 'experts'. If P is compared with *any* possible 'expert' in the simplex domain $\Delta_m = \left\{b_t : b_t \in \mathbb{R}_+^m, \sum_{i=1}^{m} b_{ti} = 1\right\}$ then (1) becomes the so-called *regret* [4]

$$r(P) = \mathbf{W}_T(P, \mathtt{B}) = \sup_{\mathbf{x}} \sup_{B \in \Delta_m} \ln \frac{W_T(B)}{W_T(P)}, \tag{2}$$

where $\sup_{B \in \Delta_m} W_T(B) = W_T^*(B)$ is the wealth achieved by BCRP. Note that P outperforms BCRP if $r(P) < 0$.

Further, P is called *universal* if it achieves asymptotically *no regret* on average for T periods and arbitrary bounded $\mathbf{x}$ with respect to BCRP [4, (1.7)]

$$\frac{1}{T} r(P) = \frac{1}{T} \ln W_T(P) - \frac{1}{T} \ln W_T^*(B) \to 0. \tag{3}$$

In the recent years there has been a growing interest and skepticism concerning the value of competitive theory to analyze on-line portfolio selection algorithms. In particular, competitive analysis is inconsistent with the more widely accepted analyses and theories based on statistical assumptions. The main criticisms are: (i) $r(P)$ gives a theoretical upper bound on the loss of P relative to BCRP but omits

to analyze its applicability in practice [2], and (ii) existing CRP based algorithms do not consider 'trading risk'. We address both drawbacks.

In short, our risk-adjusted on-line portfolio selection algorithm (RAPS) incorporates the 'trading risk' in terms of the maximum observed fluctuation of the period wealth up to time t. A systematic higher $W_T(P)$ can only be achieved by accepting a higher risk [7], i.e., a higher fluctuation. Addressing (i) we *empirically* evaluate the practical applicability of RAPS on historical NYSE data.[1] Results show that RAPS is able to beat BCRP as well as several known 'follow-the-winner' algorithms, including UP of [4]. Addressing (ii) we show that RAPS performs *provably* 'as well as' UP in the worst-case.

The rest of the paper is organized as follows. In the next section we give the necessary theoretical background. We formally present and analyze RAPS. Section 3 shows the benefits of RAPS on a numerical example. Section 4 concludes.

2 Algorithm RAPS

Without making any statistical assumptions on the nature of the stock market, [4] proves that certain P are *universal*. Cover's algorithm, UP, achieves asymptotically *no regret*.

UP: The idea is to start with the Uniform CRP (UCRP) in period $t = 1$, i.e., $b_1 = (\frac{1}{m}, \ldots, \frac{1}{m})$. For $t \geq 2$, the $\mathbf{b}$ is approximated by the past performance of the N 'experts' [4, (1.3), p. 2]

$$\hat{b}_{t+1i} = \frac{\sum_{j=1}^{N} b_{ti}^{j} \cdot W_t(B^j)}{\sum_{j=1}^{N} W_t(B^j)}, \tag{4}$$

where $W_t(B^j) = W_{t-1}(B^j) \cdot b_t^{\mathrm{T}} x_t$ denotes the *compound period wealth* of the jth 'expert' in the tth period. Thus, in hindsight, $\hat{b}_{t+1}$ is the weighted average over all 'experts' in target set B [4, p. 3].

Lemma 1 *Assume that UP competes against target set* B*. UP divides* W_o *in* N *equal parts and invests according to* B^j $(j = 1, \ldots, N)$*. Then the* terminal wealth *of UP equals* $W_T(UP) = \frac{1}{N} \sum_{j=1}^{N} W_T(B^j)$*, and its* worst-case logarithmic wealth ratio *is bounded as* [*cf.* (1)] [3, Example 10.3, p. 278]

$$\boldsymbol{W}_T(UP, \mathrm{B}) = \sup_{\boldsymbol{x}} \sup_{B^j \in \mathrm{B}} \ln \frac{W_T(B^j)}{W_T(UP)} \leq \ln N. \tag{5}$$

Lemma 2 *If* μ *is the uniform density on* Δ_m*, then UP of* [4] *satisfies* [*cf.* (2)] [3, Theorem 10.3, p. 283]

[1] http://www.cs.bme.hu/~oti/portfolio/data/nyseold.zip

$$r(UP) = \sup_{\boldsymbol{x}} \sup_{B \in \Delta_m} \ln \frac{W_T(B)}{W_T(UP)} \leq (m-1)\ln(T+1). \tag{6}$$

UP exploits the 'follow-the-winner' principle, and performs *provably* 'almost as well' as BCRP [4, Theorem 7.1].

RAPS: Let $m_t^j = \min\left\{W_o, \ldots, W_t(B^j)\right\}$ and $M_t^j = \max\left\{W_o, \ldots, W_t(B^j)\right\}$ be the minimum and maximum compound period wealth of the jth 'expert' up to time t. Then $\phi_t(B^j) = \frac{M_t^j}{m_t^j}$ equals the maximum observed fluctuation of the period wealth up to time t, and the inverse $\phi_t(B^j)^{-1}$ quantifies an experts' possible maximum *loss* up to time t. To compute $\hat{b}_{t+1}$, UP uses the experts compound period wealth. Instead, the idea of RAPS is to replace the $W_t(B^j)$ in (4) by $\phi_t(B^j)^{-1}$. Like UP, RAPS starts with UCRP in $t = 1$. For the subsequent $t \geq 2$ periods, $\mathbf{b}$ is approximated by

$$\hat{b}_{t+1i} = \frac{\sum_{j=1}^N b_{ti}^j \cdot \phi_t(B^j)^{-1}}{\sum_{j=1}^N \phi_t(B^j)^{-1}}. \tag{7}$$

Lemma 3 *Assume that all $x_{ti} \leq 1$, and that RAPS competes against a target set of $\mathtt{B}$ algorithms. RAPS divides W_o in N equal parts and invests according to B^j. Then the* terminal wealth *of RAPS equals $W_T(RAPS) = \frac{1}{N}\sum_{j=1}^N \phi_t(B^j)^{-1}$, and its* worst-case logarithmic wealth ratio *is bounded as* [*cf.* (1)]

$$\boldsymbol{W}_T(RAPS, \mathtt{B}) \leq \ln N. \tag{8}$$

Proof The proof is based on Lemma 1. We know that *iff* the assets in the portfolio do not follow a positive trend, then $x_{ti} \leq 1$. It follows $\phi_t(B^j)^{-1} = m_t^j = W_t(B^j)$ for $t = 1, \ldots, T$ and $j = 1, \ldots, N$. Thus

$$\begin{aligned}
\mathbf{W}_T(RAPS, \mathtt{B}) = \sup_{\mathbf{x}} \ln \frac{\max_{j=1,\ldots,N} \phi_T(B^j)^{-1}}{\frac{1}{N}\sum_{j=1}^N \phi_T(B^j)^{-1}} &\leq \sup_{\mathbf{x}} \ln \frac{\max_{j=1,\ldots,N} \phi_T(B^j)^{-1}}{\frac{1}{N}\max_{j=1,\ldots,N} \phi_T(B^j)^{-1}} \\
&= \sup_{\mathbf{x}} \ln \frac{\max_{j=1,\ldots,N} W_T(B^j)}{\frac{1}{N}\max_{j=1,\ldots,N} W_T(B^j)} \\
&= \ln N. \qquad \square
\end{aligned} \tag{9}$$

Under worst-case assumptions $\mathbf{W}_T(RAPS, \mathtt{B}) = \mathbf{W}_T(UP, \mathtt{B})$, cf. (5). Consequently, $r(RAPS) \leq (m-1)\ln(T+1)$ also equals UP, cf. (6). The worst-case performance of UP is basically unimprovable [3, p. 285], but UP has some practical disadvantages which are addressed by [5, 6]. We aim to answer the question if RAPS is able to outperform (some of) these algorithms from the literature in case $x_{ti} > 1$.

On-line Benchmarks: Motivated by the 'follow-the-winner' principle we limit to P which increase the b_{ti} of more successful assets. Rather than targeting BH_{best},

Table 1 Portfolio comparison in terms of the $W_T(P)$ achieved for $N = 101$

#	Assets	BCRP	BH_{best}	UP	EG (0.01)	UCRP	SCRP	RAPS	$r(RAPS)$
1	Comm. Metals and Kin Arc	**144.01**	52.02	78.47	117.15	118.69	26.36	127.96	+0.12
2	IBM and Coca Cola	15.07	13.36	14.18	15.00	15.02	5.48	**15.36**	−0.02
3	Comm. Metals and Mei Corp.	102.96	52.02	72.63	97.94	98.89	28.14	**109.57**	−0.06
4	$\bar{W}_T(P) = \frac{1}{630} \sum_{p=1}^{630} W_T(P)$	**26.57**	20.72	18.89	21.73	21.84	12.13	23.07	+0.14

these algorithms mainly track BCRP. Besides (i) UP and (ii) UCRP, we consider: (iii) Exponential Gradient (EG(η)) of [6] which aims to reduce the computational costs of UP from exponential to linear. The key parameter of EG(η) is the learning rate $\eta > 0$. For $\eta \to 0$ EG(η) reduces to UCRP [6, p. 35:11]. (iv) Successive CRP (SCRP) of [5, p. 170] which directly adopts BCRP up to the tth period, i.e., b_{t+1} equals the subsequent BCRP allocation vector (b_t^*). Note that, compared to UP and RAPS, the worst-case performance guarantees of EG(η) and SCRP are inferior (not as tight).

Offline Benchmarks: In the financial community the optimal offline benchmark is to buy-and-hold the best-performing asset of the portfolio, denoted by BH_{best} [2]. In contrast, the information theory and machine learning community considers BCRP. [4, Proposition 2.1] proved that BCRP exceeds BH_{best}. Obviously, BH_{best} and BCRP can only be computed in hindsight.

3 Numerical Results

The NYSE data set includes daily closing prices of 36 assets for 22 years ($T = 5,651$). We only consider portfolios containing $m = 2$ assets, resulting in $\binom{36}{2} = 630$ possible portfolio combinations, and limit to three pairs of assets, cf. Table 1. We selected these pairs in order to make our results comparable, cf. [4, p. 23], [6, p. 340], and [5, p. 181]. Portfolios #1 and #2 can be found in [4–6], and #3 in [4, 6]. In addition, #4 gives the average $\bar{W}_T(P)$; column $r(RAPS)$ indicates whether RAPS outperforms BCRP (< 0) or not (> 0).

#1: From [4, p. 26] we know that the outperformance of BCRP is due to the leverage effect in the posteriori computed BCRP. Thus, Cover compared UP to a randomly generated portfolio (98.4). Contrary to UP, RAPS clearly outperforms the random sample ($127.96 > 98.4$), and all P.

#2: The assets show a lockstep performance (12.21 and 13.36). Though, like UP, RAPS barely outperforms them (cf. [4, p. 23]). Further, RAPS outperforms BCRP and all P.

#3: Volatile uncorrelated assets (52.02 and 22.92) lead to great gains compared to BH_{best}. This also holds for #1 (52.02 and 4.13). RAPS clearly outperforms BCRP and all P.

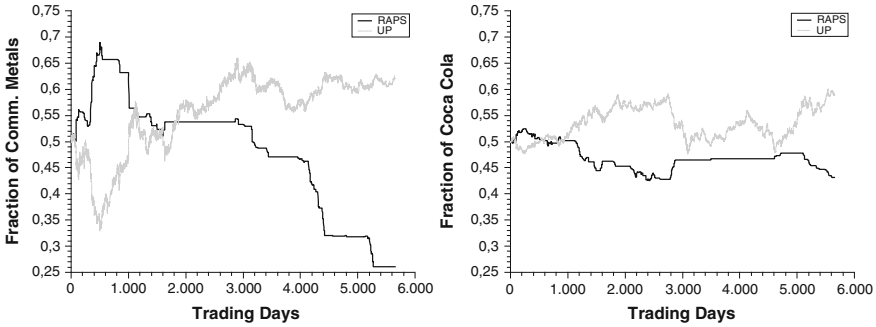


Fig. 1 Proportion of wealth ($\hat{b}_{ti}$) RAPS and UP invested in BH_{best}: #1 (*left*) and #2 (*right*)

#4: We run experiments on all 630 portfolio combinations. For each of the 36 assets $\bar{x}_{ti} > 1$ holds, where $\bar{x}_{ti} = \frac{1}{T}\sum_{t=1}^{T} x_{ti}$. On average, RAPS outperforms all P and BH_{best} but not BCRP. We claim that RAPS outperforms the online benchmarks in case the assets in the portfolio follow a *positive trend*, i.e., $\bar{x}_{ti} > 1 \; \forall \; m$ assets.

Summing up, RAPS outperforms BH_{best} and all P in all cases, and is superior to BCRP in two of three cases. Hence, Fig. 1 shows that targeting BCRP is superior to targeting BH_{best} (Comm. Metals (#1; #3) and Coca Cola (#2)) as RAPS stepwise reduces the $\hat{b}_{ti}$ invested in BH_{best}.

4 Conclusions

To the best of our knowledge, existing 'follow-the-winner' algorithms do not consider 'trading risk' when computing **b**. In contrast to existing P, RAPS targets the expert with the lowest possible loss $(\phi_t(B^j)^{-1})$. We *prove* that RAPS performs 'as well as' UP in the worst-case, and its computational costs are also exponential. Our numerical results (Table 1) are encouraging that RAPS performs well in practice. The constituent assets and all benchmark algorithms from the literature (UP, EG(0.01), UCRP, SCRP) are outperformed. In general, RAPS outperforms in case the assets in the portfolio follow a *positive trend*. Volatile uncorrelated stocks (like in #1 and #3) lead to great gains over BH_{best}. Figure 1 shows that targeting BCRP is superior to targeting BH_{best}. However, ponderous stocks (like in #2) show only modest improvements. This result is consistent with [4]. An open question is the universality of RAPS.

References

1. Borodin, A., El-Yaniv, R., & Gogan, V. (2000). On the competitive theory and practice of portfolio selection. In G. Gonnet & A. Viola (Eds.), *LATIN 2000: Theoretical informatics*. Springer, Berlin.
2. Borodin, A., El-Yaniv, R., & Gogan, V. (2004). Can we learn to beat the best stock. *Journal of Artificial Intelligence Research*, *21*(1), 579–594.

3. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. New York: Cambridge University Press.
4. Cover, T. (1991). Universal portfolios. *Mathematical Finance*, *1*(1), 1–29.
5. Gaivoronski, A. A., & Stella, F. (2000). Stochastic nonstationary optimization for finding universal portfolios. *Annals of Operations Research*, *100*, 165–188.
6. Helmbold, D., Schapire, R., Singer, Y., & Warmuth, M. (1998). On-line portfolio selection using multiplicative updates. *Mathematical Finance*, *8*(4), 325–347.
7. Modigliani, F., & Modigliani, L. (1997). Risk-adjusted performance. *Journal of Portfolio Management*, *23*(2), 45–54.

Quantified Combinatorial Optimization

Thorsten Ederer, Ulf Lorenz and Thomas Opfer

Abstract MIP and IP programming are state-of-the-art modeling techniques for computer-aided optimization. However, companies observe an increasing danger of disruptions that prevent them from acting as planned. One reason is input data being assumed as deterministic, but in reality, data is afflicted with uncertainties. Incorporating uncertainty in existing models, however, often pushes the complexity of problems that are in P or NP, to the complexity class PSPACE. Quantified integer linear programming (QIP) is a PSPACE-complete extension of the IP problem with variables being either existentially or universally quantified. With the help of QIPs, it is possible to model board-games like Gomoku as well as traditional combinatorial OR problems under uncertainty. In this paper, we present how to extend the model formulation of classical scheduling problems like the Job-Shop and Car-Sequencing problem by uncertain influences and give illustrating examples with solutions.

This research is partially supported by German Research Foundation (DFG) funded SFB 805 and by the DFG project LO 1396/2-1.

T. Ederer (✉) · T. Opfer
Discrete Optimization, TU Darmstadt, Dolivostrasse 15,
64293 Darmstadt, Germany
e-mail: ederer@mathematik.tu-darmstadt.de

T. Opfer
e-mail: opfer@mathematik.tu-darmstadt.de

U. Lorenz
Fluid Systems Technology, TU Darmstadt, Magdalenenstrasse 4,
64289 Darmstadt, Germany
e-mail: ulf.lorenz@fst.tu-darmstadt.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_17,

1 Introduction

In the past, there have been several efforts to add uncertainties into the model description of problems [1, 7]. We were involved into examinations of uncertainties in airline fleet assignment [6] and railway planning [2]. However, the ratio of implementing efforts to output was rather disappointing. Therefore, we came to the conclusion that a modeling language is needed that combines convenient MIP modeling with the ability to express uncertainties. In 2004, Subramani introduced the idea to enrich linear programs by universally quantified variables [8].

Definition 1 (*Quantified Integer Program*) Let $x = (x_1, \ldots, x_n)^T \in \mathbb{Q}^n$ be a vector with lower and upper bound vectors $l \in \mathbb{Z}^n$ and $u \in \mathbb{Z}^n$ such that $l_i \leq x_i \leq u_i$. $A \in \mathbb{Q}^{m \times n}$, $b \in \mathbb{Q}^m$ and x build a linear inequality system. Moreover, there is a vector of quantifiers $\mathcal{Q} = (\mathcal{Q}_1, \ldots, \mathcal{Q}_n)^T \in \{\forall, \exists\}^n$. Let $\mathcal{Q} \circ x \in [l, u] \cap \mathbb{Z}^n$ with the componentwise binding operator $\circ$ denote the *quantification vector* $(\mathcal{Q}_1 x_1 \in [l_1, u_1] \cap \mathbb{Z}, \ldots, \mathcal{Q}_n x_n \in [l_n, u_n] \cap \mathbb{Z})^T$ such that every quantifier $\mathcal{Q}_i$ binds the variable x_i. Let there also be a vector of objective coefficients $c \in \mathbb{Q}^n$. Each maximal consecutive subsequence of $\mathcal{Q}$ consisting of identical quantifiers is called a *quantifier block*—the corresponding ith subsequence of x is called a *variable block* B_i.

$$\text{We call} \quad z = \min_{B_1}(c^1 x^1 + \max_{B_2}(c^2 x^2 + \min_{B_3}(c^3 x^3 + \max_{B_4}(\ldots \min_{B_k} c^k x^k)))))$$
$$\mathcal{Q} \circ x \in [l, u] \cap \mathbb{Z}^n : Ax \leq b \qquad \text{(QIP)}$$

a *Quantified Integer Program with objective function* (for a minimizing existential player). Here we assume w.l.o.g. that $\mathcal{Q}_1 = \exists$ and $\mathcal{Q}_n = \exists$.

A QIP with objective can be interpreted as a two-person zero-sum game [3, 4] between a max-player who tries to make the instance infeasible or to maximize the objective function and a min-player who wants to make the instance feasible and to minimize the objective against all odds. Note that this is a short notation for a dynamic program where the players have recursively to find optimal vectors x_{B_i} with fixed $x_{B_1}, \ldots, x_{B_{i-1}}$ under consideration that the other player will set the next block of variables optimal concerning his own incentives.

2 Job Shop Scheduling

Jop Shop Scheduling is a classical optimization problem in which jobs have to be assigned to several machines such that the total time until all jobs are finished (the *makespan*) is minimized. An assignment that a job has to be processed by a machine is called a task and is given a certain duration. If some tasks depend on one another, a full or partial order of tasks can be given. Equation (2) defines the makespan. Equation (3) ensures the partial task order. The ordering indicator variables $y_{i,m,j}$ are defined by the following two equations.

Table 1 Jobshop model notation

J	Set of jobs
M	Set of machines
T	Set of tasks, $T \subseteq J \times M$
O	Taskorder, $O \subseteq T \times T$
$s_{j,m}$	Start time (integer) of task (j, m)
$d_{j,m}$	Duration of task (j, m)
$\delta_{j,m}$	Additional duration of task (j, m) in case of delay
$e_{j,m}$	Earliness of task (j, m), i.e., $e = \max\{d^1 - d^2,\ 0\}$
$\overline{e}$	Mean earliness
m	Makespan
$r_{u,j,m}$	Indicator of unary encoding of retarded task
$\tilde{r}_b$	Indicator of binary encoding retarded task
$\boldsymbol{w}$	Wrapping indicator for binary to unary translation

For solution purposes it is relevant to use as few universally quantified variables as possible. To that end, we introduce a binary encoding $\tilde{r}$ of the retardation and add existentially quantified helper variables r as unary encoding of the retardation. Equations (6) represent a linear formulation of this translation. Equations (7)–(10) are an adaption of the prior constraints for the second stage variables with uncertainly prolonged task durations. Equations (11)–(13) define the earliness caused by the existential players reaction, which is used as a penalty term for large rearrangements of the first stage planning (Table 1).

$$\min \quad m^2 + k \cdot \overline{e} + \frac{1}{M} \cdot m^1 \qquad \text{s.t.} \qquad \exists\, s^1\, y^1\, m^1\, \forall\, \tilde{r}\ \exists\, r\, w,\ s^2\, y^2\, m^2,\ e: \tag{1}$$

$$s^1_{j,m} + d_{j,m} \ \leq\ m^1 \qquad \forall\,(j,m) \in T \tag{2}$$

$$s^1_{i,m} + d_{i,m} \ \leq\ s^1_{j,n} \qquad \forall\,(i,m,j,n) \in O \tag{3}$$

$$s^1_{i,m} + d_{i,m} \ \leq\ s^1_{j,m} + M \cdot (1 - y^1_{i,m,j}) \qquad \forall\,(i,m) \in T,\ (j,m) \in T \tag{4}$$

$$s^1_{j,m} + d_{j,m} \ \leq\ s^1_{i,m} + M \cdot y^1_{i,m,j} \qquad \forall\,(i,m) \in T,\ (j,m) \in T \tag{5}$$

$$\sum_{(u,j,m) \in U} u \cdot r_{j,m} \ =\ \sum_{b \in B} 2^b \cdot \tilde{r}_b - |T| \cdot \boldsymbol{w} \quad \wedge \quad \sum_{\substack{u \in U \\ (j,m) = T_u}} r_{j,m} \ \leq\ 1 \tag{6}$$

$$s^2_{j,m} + d_{j,m} + \delta_{j,m} \cdot r_{j,m} \ \leq\ m^2 \qquad \forall\,(j,m) \in T \tag{7}$$

$$s^2_{i,m} + d_{i,m} + \delta_{i,m} \cdot r_{i,m} \ \leq\ s^2_{j,n} \qquad \forall\,(i,m,j,n) \in O \tag{8}$$

$$s^2_{i,m} + d_{i,m} + \delta_{i,m} \cdot r_{i,m} \ \leq\ s^2_{j,m} + M \cdot (1 - y^2_{i,m,j}) \qquad \forall\,(i,m) \in T,\ (j,m) \in T \tag{9}$$

$$s^2_{j,m} + d_{j,m} + \delta_{i,m} \cdot r_{i,m} \ \leq\ s^2_{i,m} + M \cdot y^2_{i,m,j} \qquad \forall\,(i,m) \in T,\ (j,m) \in T \tag{10}$$

$$e_{i,m} \ \geq\ s^1_{i,m} - s^2_{i,m} \qquad \forall\,(i,m) \in T \tag{11}$$

Table 2 Jobshop tasks

Job	Machine	Duration	Extra
Paper1	Blue	45	5
Paper1	Yellow	10	0
Paper2	Blue	20	5
Paper2	Green	10	10
Paper2	Yellow	34	0
Paper3	Blue	12	0
Paper3	Green	17	0
Paper3	Yellow	28	20

Table 3 Jobshop order

Prior task		Later task	
Paper1	Blue	Paper1	Yellow
Paper2	Green	Paper2	Blue
Paper2	Blue	Paper2	Yellow
Paper3	Yellow	Paper3	Blue
Paper3	Blue	Paper3	Green

Table 4 Solution of the jobshop example

sc.	Start times								m.s.
	1B	1Y	2B	2G	2Y	3Y	3B	3G	
	First stage solution								
	0	45	0	45	65	0	70	82	99
1	0	50	0	50	70	0	70	82	104
2	0	45	0	45	65	0	65	82	99
3	0	45	0	45	70	0	70	82	104
4	0	45	0	45	65	0	65	82	99
5	0	45	0	45	65	0	65	82	99
6	0	45	0	45	65	0	65	82	99
7	0	45	0	45	65	0	70	82	99
8	0	48	0	50	70	0	75	87	104

$$e_{i,m} \geq 0 \tag{12}$$

$$\overline{e} = \frac{1}{|T|} \cdot \sum_{(i,m) \in T} e_{i,m} \tag{13}$$

An example[1] is given in Tables 2 and 3. Table 4 depicts an optimal solution. The first-stage scheduling has the property that the planner can find a rescheduling to each possible redardation such that the worst-case makespan is minimized.

[1] The model formulation and example data are adapted from the work of Jeffrey Kantor, Christelle Gueret, Christian Prins and Marc Sevaux, cf. http://estm60203.blogspot.com/.

3 Car Sequencing

In flexible manufacturing systems, varying models of same basic product are produced. They usually require different processing times, so sequences which alternately produce different models are preferable. We consider so called $r_k : s_k$ sequencing rules [5] that restrict too frequent production of work intensive models at certain stations, that is, option k may only be produced at most r_k times per each s_k successively sequenced models.

We add uncertainty to this problem by incorporating a malfunction in the production process. It may happen that a car cannot be processed in the prescheduled order and has to be reinserted a few timesteps later after the malfunction has been corrected. The uncertainty is given by a tuple (t, t'), $t < t'$ with the new timestep t' at which the model originally scheduled at t will be processed. The cars inbetween will each be processed one timestep earlier. The resulting schedule is modeled by stage two ($s = 2$) variables—note that they may be chosen differently for each possible malfunction. The planer may react to this uncertainty by rescheduling yet another model, i.e., he chooses a tuple (u, u'), $t' < u < u'$ such that the car which was originally scheduled at u will be processed at u'. This final reschedule is given by stage three variables.

The first three equations and the first stage variables ($s = 1$) give a formulation of the original problem without uncertainty. Equation (15) ensures that for each class c the produced amount equals the given demand D_c. Equation (16) specifies that exactly one unit is produced in each time step. The $r_k : s_k$ sequencing rules are not strictly enforced—instead violations are counted by the indicator variables y^s_{k,t_0} in Eq. (17). Similar to the job shop model, we introduce a binary encoding $\tilde{m}$ and helper variables m for the uncertain machine malfunctions. Given unary encodings of the malfunction m_u and the answer a_u, we can encode the change of schedule with constraints similar to Eq. (19) (which model which parts of the schedule do *not* change) and further constraints which ensure that the cars are processed in the correct new order. These equations are rather long but not very insightful, so we skip them here.

$$\min \sum_{k \in O} \sum_{t_0 \in T^k} y^1_{k,t_0} \quad \text{s.t.} \quad \exists\, x^1\, y^1\, \forall\, \tilde{m}\ \exists\, m\, \boldsymbol{w}\,,\ x^2\, y^2\,,\ a\,,\ x^3\, y^3 : \tag{14}$$

$$\sum_{t \in T} x^s_{t,c} = D_c \qquad \forall\, c \in C\,,\ s \in S \tag{15}$$

$$\sum_{c \in C} x^s_{t,c} = 1 \qquad \forall\, t \in T\,,\ s \in S \tag{16}$$

$$\sum_{t=t_0}^{t_0+s_k} \sum_{c \in C} A_{k,c} \cdot x^s_{t,c} \leq r_k + M \cdot y^s_{k,t_0} \qquad \forall\, k \in O\,,\ t_0 \in T^k\,,\ s \in S \tag{17}$$

$$\sum_{u \in U} u \cdot m_u = \sum_{b \in B} 2^b \cdot \tilde{m}_b - |F| \cdot \boldsymbol{w} \quad \wedge \quad \sum_{u \in U} m_u \leq 1 \tag{18}$$

Table 5 Car instance

Opt	r	s	
	1	1	2
	2	2	3
Class	Cars	Opt 1	Opt 2
0	2	0	0
1	3	0	1
2	1	1	0
3	4	1	1

Table 6 Notation of the car sequencing model

O	Set of options
C	Set of classes, $C \subseteq \mathscr{P}(O)$
T	Set of timesteps ($\|T\|$ equals number of models)
T^k	Set of intervals (by first timestep) for option k , $T^k = \{1, \ldots, \|T\| - s_k + 1\}$
$r_k : s_k$	At most r_k out of s_k successively sequenced models may require option k
D_c	Demand of models of class c
$A_{k,c}$	Indicator, if models of class c require option k
$x_{t,c}$	Indicator, if a model of class c is produced at timestep t
y_{k,t_0}	Indicator, if the sequencing rule $r_k : s_k$ beginning at timestep t_0 is satisfied
S	Set of stages (duplicates of the original variables after each move)
σ	Stage index: 1 pre-scheduling, 2 state after delay, 3 re-scheduling
F	Ordered set of possible delays, $F = \{(t, t') \in T \times T,\ t < t'\}$
U	Unary encoding vector of F
d_u	Indicator of unary encoding for the delay from timestep t to t', $(t, t') \in U$
B	Binary encoding of possible delays
$\tilde{d}_b$	Indicator of binary encoding for the delay, $b \in B$
$\boldsymbol{w}$	Wrapping indicator for binary to unary translation

$$|x^2_{t,c} - x^1_{t,c}| \leq \sum_{u \in U, (t_i, t_j) = F_u, t_i < t < t_j} m_u \quad \forall c \in C\ ,\ t \in T \tag{19}$$

further stage-connecting constraints . . . (20)

An example is given by Tables 5 and 7 shows an optimal solution. The first-stage solution has the property, that the production planner can respond (column 3) to each possible malfunction (column 2) such that the second-stage production sequence has a worst-case minimal number of violated sequencing constraints (Table 6).

Table 7 Solution of the car sequencing example

Scenario	Mal.	Ans.	Production sequence
First stage solution			1, 3, 0, 3, 0, 3, 1, 2, 1, 3
1	–	(4, 5)	1, 3, 0, 3, 3, 0, 1, 2, 1, 3
2	(0, 1)	(2, 7)	3, 1, 3, 0, 3, 1, 2, 0, 1, 3
3	(0, 2)	–	3, 0, 1, 3, 0, 3, 1, 2, 1, 3
4	(0, 3)	–	3, 0, 3, 1, 0, 3, 1, 2, 1, 3
5	(0, 4)	–	3, 0, 3, 0, 1, 3, 1, 2, 1, 3
6	(0, 5)	(6, 8)	3, 0, 3, 0, 3, 1, 2, 1, 1, 3
7	(0, 6)	(7, 8)	3, 0, 3, 0, 3, 1, 1, 1, 2, 3
8	(0, 7)	(8, 9)	3, 0, 3, 0, 3, 1, 2, 1, 3, 1
9	(0, 8)	–	3, 0, 3, 0, 3, 1, 2, 1, 1, 3
10	(0, 9)	–	3, 0, 3, 0, 3, 1, 2, 1, 3, 1
11	(1, 2)	–	1, 0, 3, 3, 0, 3, 1, 2, 1, 3
12	(1, 3)	–	1, 0, 3, 3, 0, 3, 1, 2, 1, 3
13	(1, 4)	–	1, 0, 3, 0, 3, 3, 1, 2, 1, 3
14	(1, 5)	–	1, 0, 3, 0, 3, 3, 1, 2, 1, 3
15	(1, 6)	–	1, 0, 3, 0, 3, 1, 3, 2, 1, 3
16	(1, 7)	–	1, 0, 3, 0, 3, 1, 2, 3, 1, 3
17	(1, 8)	–	1, 0, 3, 0, 3, 1, 2, 1, 3, 3
18	(1, 9)	–	1, 0, 3, 0, 3, 1, 2, 1, 3, 3
19	(2, 3)	(4, 9)	1, 3, 3, 0, 3, 1, 2, 1, 3, 0
20	(2, 4)	–	1, 3, 3, 0, 0, 3, 1, 2, 1, 3
21	(2, 5)	–	1, 3, 3, 0, 3, 0, 1, 2, 1, 3
22	(2, 6)	–	1, 3, 3, 0, 3, 1, 0, 2, 1, 3
23	(2, 7)	–	1, 3, 3, 0, 3, 1, 2, 0, 1, 3
24	(2, 8)	–	1, 3, 3, 0, 3, 1, 2, 1, 0, 3
25	(2, 9)	–	1, 3, 3, 0, 3, 1, 2, 1, 3, 0
26	(3, 4)	–	1, 3, 0, 0, 3, 3, 1, 2, 1, 3
27	(3, 5)	–	1, 3, 0, 0, 3, 3, 1, 2, 1, 3
28	(3, 6)	–	1, 3, 0, 0, 3, 1, 3, 2, 1, 3
29	(3, 7)	–	1, 3, 0, 0, 3, 1, 2, 3, 1, 3
30	(3, 8)	–	1, 3, 0, 0, 3, 1, 2, 1, 3, 3
31	(3, 9)	–	1, 3, 0, 0, 3, 1, 2, 1, 3, 3
32	(4, 5)	–	1, 3, 0, 3, 3, 0, 1, 2, 1, 3
33	(4, 6)	–	1, 3, 0, 3, 3, 1, 0, 2, 1, 3
34	(4, 7)	(8, 9)	1, 3, 0, 3, 3, 1, 2, 0, 3, 1
35	(4, 8)	–	1, 3, 0, 3, 3, 1, 2, 1, 0, 3
36	(4, 9)	–	1, 3, 0, 3, 3, 1, 2, 1, 3, 0
37	(5, 6)	(7, 9)	1, 3, 0, 3, 0, 1, 3, 1, 3, 2
38	(5, 7)	–	1, 3, 0, 3, 0, 1, 2, 3, 1, 3
39	(5, 8)	–	1, 3, 0, 3, 0, 1, 2, 1, 3, 3
40	(5, 9)	–	1, 3, 0, 3, 0, 1, 2, 1, 3, 3
41	(6, 7)	(8, 9)	1, 3, 0, 3, 0, 3, 2, 1, 3, 1
42	(6, 8)	–	1, 3, 0, 3, 0, 3, 2, 1, 1, 3
43	(6, 9)	–	1, 3, 0, 3, 0, 3, 2, 1, 3, 1
44	(7, 8)	–	1, 3, 0, 3, 0, 3, 1, 1, 2, 3
45	(7, 9)	–	1, 3, 0, 3, 0, 3, 1, 1, 3, 2
46	(8, 9)	–	1, 3, 0, 3, 0, 3, 1, 2, 3, 1

4 Conclusion

We presented Quantified Integer Programming as an intuitive modelling language to generate recoverable robust solutions for classical scheduling problems that were extended by various uncertain influences.

References

1. Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009) *Robust Optimization*. Princeton University.
2. Berger, A., Hoffmann, R., Lorenz, U., & Stiller, S. (2011). Online railway delay management: Hardness, simulation and computation. *Simulation*, *87*(7), 616–629.
3. Ederer, T., Lorenz, U., Martin, A., Wolf, J. (2011) *Quantified Linear Programs: A Computational Study. Algorithms-ESA 2011* (pp. 203–214). Berlin: Springer.
4. Ederer, T., Lorenz, U., Opfer, T., & Wolf, J. (2011). *Modelling Games with the help of Quantified Integer Linear Programs. ACG 13*. Berlin: Springer.
5. Fliedner, M., & Boysen, N. (2008). Solving the car sequencing problem via branch & bound. *European Journal of Operational Research*, *191*(3), 1023–1042.
6. Grothklags S., Lorenz U., & Monien B. (2009). From state-of-the-art static fleet assignment to flexible stochastic planning of the future. *Algorithmics of large and complex networks* (pp. 140–165).
7. Liebchen, C., Lübbecke, M. E., Möhring, R. H., & Stiller, S. (2009). The concept of recoverable robustness, linear programming recovery, and railway applications. *Robust and online large-scale optimization* 1–27.
8. Subramani, K. (2004). *Analyzing selected quantified integer programs*. Berlin: Springer.

On the Modeling of Recharging Stops in Context of Vehicle Routing Problems

Stefan Frank, Henning Preis and Karl Nachtigall

Abstract Caused by regulations regarding to climate protection, battery electric vehicles (BEVs) offer great opportunities in context of ecological compatibility of urban transport systems. Therefore, operating models in context of vehicle routing are required. Because of the BEVs more restrictive driving range in comparison to vehicles with an internal combustion engine (ICE) and due to the fact of a less-developed network of service stations, model formulations have to include the possibility of recharging at dedicated locations. So additional restrictions in formulations are needed to handle the maximum range depending on battery capacity. There were published only a small number of articles addressed to energy consumption, battery range and possible recharging stops in mixed-integer programming (MIP) formulations in the underlying practice relevant Vehicle Routing Problem with Time Windows (VRPTW) over the past few years. So we describe different MIP-formulations for an enhanced VRPTW, considering capacity restrictions concerning to cargo and energy, customer time windows and the capability of charging stops. Effects of these formulations are shown for small-sized problems, while a column generation approach is presented for more realistic problem instances.

1 Introduction

In terms of including a distance range caused by battery capacity of BEVs and also charging stations to recharge batteries due to tour examining in an underlying VRPTW, Erdogan and Miller-Hooks [4], Preis et al. [5] and Schneider et al. [6] give

S. Frank (✉) · H. Preis · K. Nachtigall
Institute of Logistics and Aviation, Technische Universität Dresden, 01062 Dresden, Germany
e-mail: stefan.frank@tu-dresden.de

H. Preis
e-mail: henning.preis@tu-dresden.de

K. Nachtigall
e-mail: karl.nachtigall@tu-dresden.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_18,

MIP-formulations. In these works additional resource constraints are included to restrict route length up to battery capacity, and also possibilities to recharge at dedicated locations, represented as further nodes in the graph. These nodes are implemented as so called dummy sets, which means that any of the nodes representing charging stations is included several times in the graph. The reason therefore is to allow several visits at these charging stations and the so needed timestamps of service beginning, which could be allocated only once. Following this, it might be tough defining the number of dummies representing a charging station. For large problems there are multiple possibilities of including them in tours. A slightly generous set unnecessarily forces degeneration and increases inherent the solving time. To the best of our knowledge, there are no formulations published which include charging stations without defining dummy sets of them so far. Hence, following the problem definition and notation we illustrate a standard formulation and a possible model without the need of dummy sets.

Let $G = (V, E, c)$ be the tuple on the complete directed graph G with vertex set V, arc set $E = V \times V$, and cost c, here represented by distances. Let V be denoted by $V = D \cup K \cup L$ with depot $D = \{0\}$, the set of n customers $K = \{1, 2, \ldots, n\}$, and the set of p discrete located charging stations $L = \{n+1, n+2, \ldots, n+p\}$. All customers are associated with a non-negative demand b_i and service time windows $\left[t_i^b, t_i^e\right]$. Traveltimes for arcs are represented by t_{ij}^F, including the service times t_i^S. With each arc we also associate cost for energy consumption of the empty vehicle c_{ij}^F and a component c_{ij}^L depending on the payload. Furthermore we define variables t_i of beginning service at vertex i and variables e_i as the remaining energy level on arrival at vertex i. The cargo capacity of each vehicle from the homogenous fleet is given by C and the battery capacity by B. The maximum duration until the arrival at the depot is denoted by T. Further decision and arc flow variables are defined in the sections. Conveniently, we identify the following models as Vehicle Routing Problem with Charging Stations (VRPCS).

2 A Two-Index Formulation

This formulation involves decision variables x_{ij} equal to 1 if arc $(i, j) \in E$ with $i \neq j$ is used by a vehicle and 0 otherwise, and variables m_{ij} representing the allocated loading weight. Each discrete located charging station in set L needs to be defined multiple times. Thus, set L is expanded to

$$L = \left\{n+1, \ldots, n+q_1, n+q_1+1, \ldots, n+q_1+q_2, \ldots, n+q_1+\cdots+q_p\right\}$$

where q_i is the given number of dummies of a charging station i, and p is the given number of different located stations.

$$\text{VRPCS1} \quad \min \sum_{i \in V} \sum_{j \in V} c_{ij}\, x_{ij} \tag{1}$$

subject to

$$\sum_{i \in V} x_{ij} = 1 \quad \forall j \in K \tag{2}$$

$$\sum_{i \in V} x_{ij} \leq 1 \quad \forall j \in L \tag{3}$$

$$\sum_{i \in V} x_{ij} - \sum_{i \in V} x_{ji} = 0 \quad \forall j \in V \tag{4}$$

$$\sum_{i \in V} m_{ij} - \sum_{i \in V} m_{ji} = b_j \quad \forall j \in V \backslash D \tag{5}$$

$$0 \leq m_{ij} \leq C x_{ij} \quad \forall i, j \in V \tag{6}$$

$$t_i \leq t_j - t_{ij}^F x_{ij} + T\left(1 - x_{ij}\right) \quad \forall i \in V; j \in V \backslash D \tag{7}$$

$$t_i + t_{i0}^F x_{i0} \leq T \quad \forall i \in V \backslash D \tag{8}$$

$$t_i^b \leq t_i \leq t_i^e \quad \forall i \in K \tag{9}$$

$$e_j \leq e_i - c_{ij}^F x_{ij} - c_{ij}^L m_{ij} + B\left(1 - x_{ij}\right) \quad \forall i \in K; j \in V \tag{10}$$

$$e_j \leq B - c_{ij}^F x_{ij} - c_{ij}^L m_{ij} \quad \forall i \in V \backslash K; j \in V \tag{11}$$

$$0 \leq e_i \leq B \quad \forall i \in V \tag{12}$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \in V \tag{13}$$

Constraints (2) and (3) ensure that each customer is served exactly once, while charging stations may be visited once. Constraints (4) and (5) impose flow conservation at all vertices. Constraints (6) force that only used arcs are allocated with payload considering cargo capacity. Constraints (7)–(9) guarantee the compliance of service time windows, maximum route duration and furthermore the prevention of subtours. Constraints (10)–(12) define feasible energy levels, which includes recharging at charging stations and restricts possible recuperation up to battery capacity.

3 A Mixed Two/Three-Index Formulation

This formulation involves decision variables x_{ik} equal to 1 if arc $(i, k) \in E$ is used by a vehicle and 0 otherwise, while i and k both are not vertices from the set of charging stations with $i \neq k$, and decision variables y_{ijk} equal to 1 if a vehicle traverses a charging-station j starting from node i and ending at node k and 0 otherwise, while

i and k are not nodes from the set of charging stations and consequently $i \neq k$. Besides, we accordingly involve variables m_{ik} and n_{ijk} representing the allocated payload of a vehicle on an arc, respectively path.

$$\text{VRPCS2} \quad \min \sum_{i \in V \setminus L} \sum_{j \in L} \sum_{k \in V \setminus L} \left(c_{ik}\, x_{ik} + \left(c_{ij} + c_{jk} \right) y_{ijk} \right) \tag{14}$$

subject to

$$\sum_{i \in V \setminus L} x_{ik} + \sum_{i \in V \setminus L} \sum_{j \in L} y_{ijk} = 1 \quad \forall k \in K \tag{15}$$

$$\sum_{i \in V \setminus L} x_{ik} + \sum_{i \in V \setminus L} \sum_{j \in L} y_{ijk} - \sum_{i \in V \setminus L} x_{ki} - \sum_{i \in V \setminus L} \sum_{j \in L} y_{kji} = 0 \quad \forall k \in V \setminus L \tag{16}$$

$$\sum_{i \in V \setminus L} m_{ik} + \sum_{i \in V \setminus L} \sum_{j \in L} n_{ijk} - \sum_{i \in V \setminus L} m_{ki} - \sum_{i \in V \setminus L} \sum_{j \in L} n_{kji} = b_k \quad \forall k \in K \tag{17}$$

$$0 \leq m_{ik} \leq C x_{ik} \quad \forall i, k \in V \setminus L \tag{18}$$

$$0 \leq n_{ijk} \leq C y_{ijk} \quad \forall i, k \in V \setminus L;\, j \in L \tag{19}$$

$$t_k \geq t_i + t_{ik}^F x_{ik} + \left(t_{ij}^F + t_{jk}^F \right) y_{ijk} - T \left(1 - x_{ik} - y_{ijk} \right) \quad \forall i \in V \setminus L;\, j \in L;\, k \in K \tag{20}$$

$$t_0^e \geq t_i + t_{i0}^F x_{i0} + \left(t_{ij}^F + t_{j0}^F \right) y_{ij0} - T \left(1 - x_{i0} - y_{ij0} \right) \quad \forall i \in K;\, j \in L \tag{21}$$

$$t_i^b \leq t_i \leq t_i^e \quad \forall i \in V \setminus L \tag{22}$$

$$e_k \leq e_i - c_{ik}^F\, x_{ik} - c_{ik}^L\, m_{ik} + B \left(1 - x_{ik} \right) \quad \forall i \in K;\, k \in V \setminus L \tag{23}$$

$$e_k \leq B - c_{0k}^F\, x_{0k} - c_{0k}^L\, m_{0k} \quad \forall k \in K \tag{24}$$

$$e_i \geq c_{ij}^F\, y_{ijk} + c_{ij}^L\, n_{ijk} \quad \forall i \in K;\, j \in L;\, k \in V \setminus L \tag{25}$$

$$c_{0j}^F\, y_{0jk} + c_{0j}^L\, n_{0jk} \leq B \quad \forall j \in L;\, k \in V \setminus L \tag{26}$$

$$e_k \leq B - c_{jk}^F\, y_{ijk} - c_{jk}^L\, n_{ijk} \quad \forall i, k \in V \setminus L;\, j \in L \tag{27}$$

$$0 \leq e_i \leq B \quad \forall i \in V \setminus L \tag{28}$$

$$x_{ik} \in \{0, 1\} \quad \forall i, k \in V \setminus L \tag{29}$$

$$y_{ijk} \in \{0, 1\} \quad \forall i, k \in V \setminus L;\, j \in L \tag{30}$$

Because of (15) a customer must be reached only once, either directly from the depot or another customer or similarly via a charging station. So because of (16) and (17) for flow conservation and valid payloads, the connectivity is given. Constraints (18) and (19) enforce that used arcs, respectively traversing paths, could be allocated with payload, which does not exceed the vehicles cargo capacity. Following constraints (20)–(22), feasible time stamps, route duration and subtour prevention is guaranteed. Constraints (23)–(28) impose feasible energy levels of batteries.

4 A Set-Partitioning Formulation

The VRPCS can also be reformulated as set-partitioning problem. Therefore, let R be the set of feasible routes, each starting and ending at the depot, visiting several customers once and charging stations maybe repeatedly. Coefficients δ_{ir} takes value 1 if customer i is visited in route r and 0 otherwise. The cost of a route r represented by its distance is denoted by λ_r, while variables γ_r takes value 1 if route r is part of the solution and 0 otherwise.

$$\text{VRPCS3 } MP \quad \min \sum_{r \in R} \lambda_r \, \gamma_r \tag{31}$$

subject to

$$\sum_{r \in R} \delta_{ir} \, \gamma_r = 1 \quad \forall i \in K \tag{32}$$

$$\gamma_r \in \{0, 1\} \quad \forall r \in R \tag{33}$$

Constraints (32) ensure that each customer takes part in one of the selected routes. Because it is not comprehensive to generate all possible routes of R in the master-problem MP, a column generation approach usually is used to add feasible routes. Therefore, the dual variables π_i from the MP, representing the marginal cost of customer $i \in K$, are attached in the following subproblem SP. Additional variables z_k equal to 1 if customer k is part of a generated route or 0 otherwise are included to ensure flow conservation.

$$\text{VRPCS3 } SP \quad \min \sum_{i \in V \setminus L} \sum_{j \in L} \sum_{k \in V \setminus L} \left((c_{ik} - \pi_i) \, x_{ik} + \left(c_{ij} + c_{jk} - \pi_i \right) y_{ijk} \right) \tag{34}$$

subject to

$$\sum_{k \in K} x_{0k} + \sum_{j \in L} \sum_{k \in K} y_{0jk} = 1 \tag{35}$$

$$\sum_{i \in V \setminus L} x_{ik} + \sum_{i \in V \setminus L} \sum_{j \in L} y_{ijk} - \sum_{i \in V \setminus L} x_{ki} - \sum_{i \in V \setminus L} \sum_{j \in L} y_{kji} = 0 \quad \forall k \in V \setminus L \tag{36}$$

$$\sum_{i \in V \setminus L} m_{ik} + \sum_{i \in V \setminus L} \sum_{j \in L} n_{ijk} - \sum_{i \in V \setminus L} m_{ki} - \sum_{i \in V \setminus L} \sum_{j \in L} n_{kji} = b_k z_k \quad \forall k \in K \quad (37)$$

$$\sum_{i \in V \setminus L} x_{ik} + \sum_{i \in V \setminus L} \sum_{j \in L} y_{ijk} = z_k \quad \forall k \in K \quad (38)$$

$$z_k \in \{0, 1\} \quad \forall k \in K \quad (39)$$

and constraint sets (18)–(30).

Due to constraint (35) exactly one customer is served either directly from the depot or with the detour via a charging station. Constraints (36) ensure connectivity. Constraints (37) guarantee flow conservation for allocated customers, which is enforced by constraints (38).

For fundamentals to column generation and approaches depending on the VRPTW we refer to Desrochers et al. [3], Toth and Vigo [8] and Desaulniers et al. [2]. For the sake of brevity we outline several acceleration techniques of our so far approach. To get good dual values at the beginning we use several construction heuristics and local search for initial routes in *MP*. Branching in *MP* occurs on the decision if customers are allocated to the same or to different vehicles. Constraints depending on this branching are added to *SP*. If there exists more than one solution of *SP* with reduced cost less than zero, not only the best route is added to *MP*, but only the best of different solutions with same customer allocation. To accelerate the solving process we also use several construction heuristics and local search in *SP* to find feasible routes. Further work will aim at solving *SP* with the help of a labeling algorithm.

Table 1 Average results

Customers	Vehicles	Stops	Time ratio
10	2.44	2.24	3.47
15	2.78	2.61	3.95
20	3.21	2.89	5.99
25	4.18	3.42	21.06

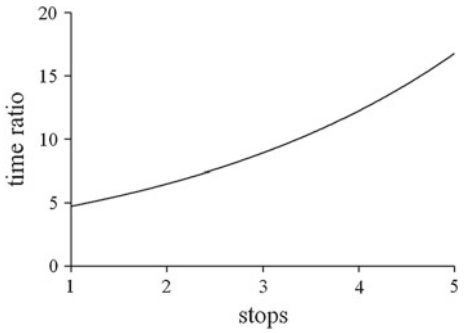


Fig. 1 Time ratio dependent on stops

5 Preliminary Results

Experiments were made with SCIP version 3.0.1 [1] on academic instances and modified versions of the well known instances from Solomon [7], all with three charging stations and also three dummies in VRPCS1. Table 1 shows the average results over 100 instances each for 10 to 25 customers. The number of needed vehicles, the number of realized charging stops and the ratio of solving time for VRPCS1 compared to VRPCS2 dependent on the number of customers are stated. Figure 1 shows the average ratio of solving time for VRPCS1 in relation to VRPCS2 over all tested instances dependent on charging stops. It could be seen that the solving time ratio raises with an increasing number of customers respectively charging stops, so the model formulation without the use of dummy sets has a positive impact.

References

1. Achterberg, T. (2009). SCIP: solving constraint integer programs. *Mathematical Programming Computation*, *1*(1), 1–41.
2. Desaulniers, G., Desrosiers, J., & Solomon, M. M. (2005). *Column generation*. New York: Springer.
3. Desrochers, M., Desrosiers, J., & Solomon, M. M. (1992). A new optimization algorithm for the vehicle routing problem with time windows. *Operations Research*, *40*, 324–354.
4. Erdogan, S., & Miller-Hooks, E. (2012). A green vehicle routing problem. *Transportation Research Part E*, *48*, 100–114.
5. Preis, H., Frank, S., & Nachtigall, K. (2014). Energy-optimized routing of electric vehicles in urban delivery systems. In: S. Helber, M. Breitner, D. Rösch, C. Schön, J. -M. Graf von der Schulenburg, P. Sibbertsen, M. Steinbach, S. Weber, & A. Wolter (Eds.), *Operations research proceedings 2012* (pp. 583–588). Berlin: Springer
6. Schneider, M., Stenger, A., & Goeke, D. (2012). The electric vehicle routing problem with time windows and recharging stations. Technical report 02/2012
7. Solomon, M. M. (1987). Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, *35*, 254–265.
8. Toth, P., & Vigo, D. (2002). *The vehicle routing problem*. Philadelphia: Siam.

Demand Fulfillment in an Assemble-to-Order Production System

Sebastian Geier and Bernhard Fleischmann

Abstract We consider a computer manufacturer who assembles customized final products from various components. Customer orders specify the product configuration, the quantity and a desired delivery date. The online order promising (OP) process must announce a first promised delivery date to the customer. Demand fulfillment in this Assemble-to-Order (ATO) case is still little investigated and differs remarkably from the more popular Make-to-Stock (MTS) case: Bottlenecks are the assembly capacity and the stocks of components, which are available to promise (ATP). An important task of the demand fulfillment, besides OP, is Demand Supply Matching (DSM), i.e. deciding on the assembly date of orders and eventually changing the delivery date of promised orders (repromising). We present a new concept for demand fulfillment in the ATO case which consists of online OP for single orders arriving during the day and DSM once a day, linked in a rolling-horizon scheme. The DSM is based on a mixed integer programming (MIP) model which simultaneously determines assembly and delivery dates for all promised orders. We report on a case study with real data of a computer manufacturer with more than 10,000 orders on hand and 2,000 different components.

1 Introduction

Taking care of the fulfillment of customer orders is of great practical importance for most companies, but this task has been neglected in the research on supply

S. Geier (✉)
University of Augsburg, Sustainable Operations and Logistics, Universitaetsstrasse 16, 86135 Augsburg, Germany
e-mail: sebastian-geier@gmx.de

B. Fleischmann
University of Augsburg, Production and Supply Chain Management, Universitaetsstrasse 16, 86135 Augsburg, Germany
e-mail: bernhard.fleischmann@wiwi.uni-augsburg.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_19,

chain planning in the past. In the last few years the demand fulfillment gained more attention, probably due to the rise of online retailing and EDI-connection in business-to-business relationships. We consider the case where customer orders arrive during the day and every order has to be confirmed instantly. This task is referred to as online order promising (OP) and leads to a first promised delivery date. Further tasks are necessary to fulfill the order completely, depending on the decoupling point of the underlying production system. In MTS production customer orders are fulfilled from finished goods on stock and planned production, which constitute the quantities available to promise (ATP). However, in ATO production systems, such as computer manufacturing, customer orders initiate the final assembly of (customized) products (e.g. PCs or Servers) from various components (like casing, mainboard, HDD, GPU, RAM, optical drives, etc.). Thus bottlenecks are the assembly capacity and the ATP quantities of components. Typically the components have a n:m-relationship to the orders, i.e. a specific order requires several different components and a specific component can occur in several different orders. The promised delivery dates are simultaneously based on the ATP-quantities of components and on capacity. Due to the assembly step, the order fulfillment time for ATO is longer than for MTS. During this time, after the first OP, unforeseen events like faulty material supply, machine breakdowns or the arrival of urgent new orders can happen As a consequence, it may become impossible to meet all promised delivery dates. Hence a very important task of demand fulfillment in ATO-production is to monitor the promised delivery dates during the order fulfillment time. We refer to this task as short term Demand Supply Matching (DSM). DSM decides on the assembly dates of orders and eventually on repromising, i.e. changing the delivery dates of some promised orders. In the worst case, repromising leads to a cancellation of promised orders.

The next section explains a rolling horizon planning concept for online OP and DSM. Section 3 develops a MIP model for the DSM. Section 4 shows some computational results of the rolling horizon procedure for a real life case study from an international computer manufacturer.

2 Rolling Horizon Planning for Demand Fulfillment

Figure 1 shows the planning scheme for OP and DSM in a rolling horizon. OP takes place for every single order at the arrival and entails an update of the ATP quantities of the concerned components. By contrast, DSM runs once a day, overnight, for the whole set of promised, but yet unfulfilled orders. It respects the interdependency of these orders, which compete for common components and for capacity. Thus, repromising may improve the delivery dates for some important orders at the expense of other orders. The new delivery dates are determined by reserving ATP quantities and capacity for every order on appropriate days of the planning horizon. As a result, the remaining free ATP quantities are a starting point for the OP at the next day. More details about the use of ATP quantities are explained in [1].

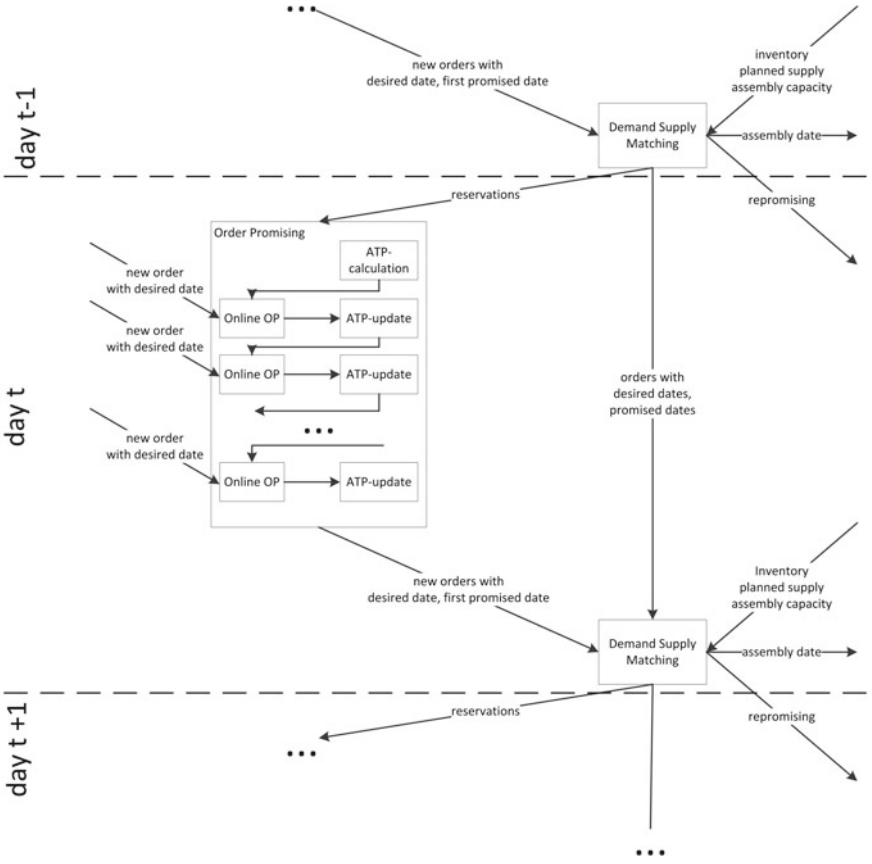


Fig. 1 Rolling horizon planning for demand fulfillment

3 Mixed-Integer Program for Demand Supply Matching

Table 1 specifies the notation for the DSM model. The main data are, day by day, the assembly capacity and the supply of every component, which consists for day 1 of the initial stock and for the days $t = 2, \ldots, T$ of known or planned inflow from suppliers or from production. The main decisions are, for every order, the delivery date, the quantities and days of assembling, and the reservation of supply of the required components. In order to avoid splitting the delivery over several dates which is not allowed, the delivery date is expressed by binary variables z_{it}. Assembling may be split and must take place on days s prior to delivery. Every assembly quantity y_{is} requires sufficient reservations x_{ijr} of all relevant components j on previous days r.

Table 1 Symbols for the demand supply matching model

t	Index of periods $t = 1, \ldots, T, T+1$
j	Index of components j = 1, . . . , J
i	Index of order i = 1, . . . , I
T	Number of periods in the planning horizon
$T+1$	Dummy-period, corresponding to non-fulfillment of orders
Order parameters	
d_i	Order quantity for order i
t_i^W	Desired delivery date for order i
r_i	Priority of order i
a_{ij}^M	Material coefficient of component j for a single unit of order i
a_i^C	Assembly capacity coefficient for a single unit of order i
c_{it}^F	Earliness and delay penalty for delivery of a single unit of order i in period t
c_i^N	Penalty costs for non-fulfillment of order i
c_j^1	Holding cost for component j
c_i^2	Holding cost for the added value of assembling one unit of order i
c_j^3	Holding cost for one unit of order i
I_i^0	Initial stock of finished products for order i
$\mathscr{L}_i$	Delivery window for order i: $\mathscr{L}_i = \left\{t_i^W - \max Earliness_i, \ldots, t_i^W, \ldots, t_i^W + \max Delay_i\right\} \cup \{T+1\}$
$\mathscr{C}_i$	Set of necessary components for order i
Global parameter	
S_{jt}	Supply of component j in period t (for t = 1: initial stock)
C_t	Assembly capacity in period t
Decision variables	
x_{ijt}	Reservation of component j for order i in period t
y_{it}	Assembly quantity for order i in period $t \in \mathscr{P}_i$
z_{it}	Delivery quantity for order i in period $t \in \mathscr{L}_i$

$$Z_{DSM} = \min \sum_i \sum_{t=1}^{T} c_{it}^F \cdot L_{it} \cdot d_i + \sum_i L_{i,T+1} \cdot d_i \cdot c_i^N \tag{1}$$

$$+ \sum_{t=1}^{T} (T-t) \left(\sum_i \sum_j c_j^1 \cdot x_{ijt} + \sum_i c_i^2 \cdot y_{it} - \sum_i c_i^3 \cdot L_{it} \cdot d_i \right) \tag{2}$$

subject to

$$\sum_i x_{ijt} \leq S_{jt} \quad \forall j, t \leq T \tag{3}$$

$$\sum_i a_i^C \cdot y_{it} \leq C_t \quad \forall t \leq T \tag{4}$$

$$\sum_{t \in \mathscr{L}_i} L_{it} = 1 \quad \forall i \tag{5}$$

$$\sum_{r=1}^{t} x_{ijr} \geq a_{ij}^{M} \cdot \sum_{s \in \mathscr{L}_i; s \leq t} y_{is} \quad \forall i, j \in \mathscr{C}_i, t \in \mathscr{L}_i \tag{6}$$

$$I_i^0 + \sum_{r \in \mathscr{L}_i; r \leq t} y_{ir} \geq \sum_{s \in \mathscr{L}_i; s \leq t} L_{is} \cdot d_i \quad \forall i, t \in \mathscr{L}_i \tag{7}$$

$$x_{ijt} \geq 0 \quad \forall i, j, t = 1, \ldots, T+1 \tag{8}$$

$$y_{it} \geq 0 \quad \forall i, t \in \mathscr{L}_i \tag{9}$$

$$L_{it} \in \{0; 1\} \quad \forall i, t \in \mathscr{L}_i \tag{10}$$

Part (1) of the objective function corresponds to the minimization of the cost of deviations from the desired delivery dates and the cost for non-fulfillment whereas part (2) minimizes the inventory costs for order specific stock of components and of finished products. Constraints (3) ensure that the reservation of components does not exceed the supply, and (4) is the capacity restriction. Constraints (5) enforce a unique delivery date, which may also be in the dummy period. Constraints (6) and (7) express, for every order, the dependencies between reservations of components, assembly quantities and delivery date, as explained before. The model formulation and the parametrization can only be sketched here, details can be found in [2].

4 Computational Results

The presented concept has been tested with several real-life data sets, provided by a European computer manufacturer, with more than 10,000 promised orders and 2,000 components. The results can be influenced by determining the penalty cost rates for unpunctual delivery and for non-fulfillment. As an example, Fig. 2 shows a contrary behavior for the two objectives minimal deviation from the desired delivery date and minimal number of non-fulfilled orders. With higher relative costs for a non-fulfillment of orders, not only the proportion of fulfilled orders increases, but also the proportion of on-time delivery decreases. Thus, the trade-off between on-time delivery and fulfillment of orders has to be considered carefully. Stability of the promised delivery dates is a further objective of great practical importance. In the test runs for the rolling-horizon scheme, additional penalty costs were introduced for repromising. We simulated DSM runs on two consecutive days, starting with the determination of the delivery dates for 10,493 orders on day $t-1$. The resulting ATP quantities were used for the online OP for 1,203 arriving orders on day t. According

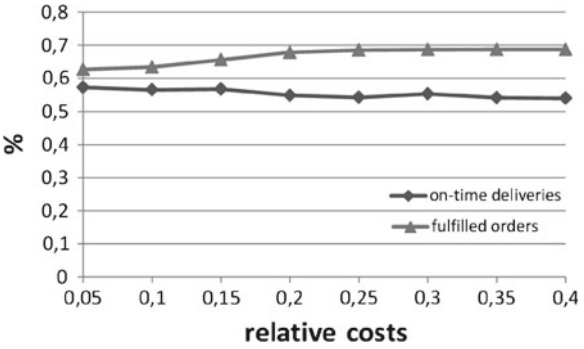


Fig. 2 Effect of increasing costs for non-fulfillment on on-time deliveries and fulfilled orders

Table 2 Resulting changes in delivery plan for rolling planning

Relative cost	Percentage of improved dates	Percentage of deteriorated dates
0.0	7.6	2.5
0.01	6.0	1.0
0.04	5.6	0.9
0.64	4.0	0.2
1.28	3.9	0.0

to the DSM planning, 1,710 orders were fulfilled on day t. Then, a DSM run for day t was performed. Table 2 shows results for different values of the repromising cost. For penalty costs of 1.28 times the order priority r_i, deviations of delivery dates, representing a deterioriation of delivery service, can be avoided, while some improvements of delivery dates (mainly of newly arrived orders on day t) are even possible.

5 Conclusions

The study shows that in the demand fulfillment for ATO production, DSM plays an important role. Combined with the online OP for single orders in a rolling horizon scheme, it generates valid and stable delivery plans. It improves the results of the OP by taking the interdependency of all orders into account. Future research effort is required for developing advanced concepts for order promising in ATO production, in particular the incorporation of customer classes.

References

1. Fleischmann, B., & Geier, S. (2011). Global available-to-promise (global ATP). In H. Stadtler, et al. (Eds.), *Advanced planning in supply chains* (pp. 195–215). Heidelberg: Springer.
2. Geier, S. (2013). Demand fulfillment bei assemble-to-order-Fertigung—Analyse, Optimierung und Anwendung in der Computer-Industrie. Dissertation, Universitaet Augsburg.

Integrating Keyword Advertising and Dynamic Pricing for an Online Market Place

Thomas Goertz, Jella Pfeiffer, Henning Schmidt and Franz Rothlauf

Abstract Keyword Advertising is a main marketing instrument for e-commerce companies in order to generate traffic from search engines on their website. The costs for Keyword Advertising are determined in an auction that is conducted for every single search query, which is entered in by a user. In case of an online market place, each adlink provided by the search engine refers to an ordered list of products on the website of the online market place. Hereby, the price of the product is oftentimes one important criteria for the user when deciding for one or the other product from the list. However, existing models assume the price of products to be exogeneous. By taking into account the prices of linked products as a further class of decision variables, we propose a joint version of the advertiser's decision problem that, apart from finding the optimal bidding strategy for Keyword Advertising, also finds the optimal pricing strategy for the offered products under a budget restriction and capacity constraints.

1 Introduction

Keyword Advertising has grown into a multibillion-dollar business. According to a survey conducted by PwC the total annual spend for Keyword Advertising in the United States in 2012 was $16.84 billion, which is 13 % increase compared to 2011. Keyword Advertising is a service offered by Internet search engines which enables

T. Goertz (✉) · J. Pfeiffer · H. Schmidt · F. Rothlauf
Johannes Gutenberg University Mainz, Jakob-Welder Weg 9, 55128 Mainz, Germany
e-mail: goertzt1@students.uni-mainz.de

J. Pfeiffer
e-mail: jella.pfeiffer@uni-mainz.de

H. Schmidt
e-mail: henning.schmidt@gmx.de

F. Rothlauf
e-mail: rothlauf@uni-mainz.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_20,

advertisers to personalize online-advertisement because they allow for displaying an ad that fits to the customer's search query. In most cases, the costs for Keyword Advertising are determined in an auction that is automatically conducted for every single search query entered by a user. The advertisers' bids for keywords determine the position of their related ad placements and the prices that they need to pay to the search engine whenever a customer actually clicks on their ads (cost-per-click). The most important Keyword Advertising service, *Google AdWords*, utilizes the Generalized Second Price Auction (GSP) to sell their slots for paid ad placements (see [2]): In case of a click on an advertiser's adlink, displayed on position i, the advertiser has to pay the submitted bid of the advertiser on position $i + 1$. In previous work, several authors have shown that bidding the true value of a keyword in a GSP is not optimal (see [2, 3]). In addition to the latter finding, multiple empirical studies in recent years have shown that top positions lead to more clicks on the advertiser's ad placement but also at higher costs [4, 5]. As shown by Ghose and Yang [5], the ad position does also affect the customer's purchase decision, i.e., better positions lead to higher purchase probabilities. Hence, strategically bidding is a necessary precondition in order to maximize an advertiser's revenue from Keyword Advertising.

Current Keyword Advertising models rely on a calculation of revenue that assume the value per keyword as fixed (see [6]). This approach, however, does not take into account that an advertiser has the possibility to balance higher costs for better ad positions with increased prices of the advertised products. Furthermore, existing models ignore the impact of limited capacity of each product on the optimal bidding policy. Especially in cases where the total expected demand exceeds the available capacity of requested products, this may lead to wrong decisions.

In this work we present an extended Keyword Advertising model that integrates prices of the advertised products as additional decision variables because we argue that this approach has the potential to further boost an advertiser's revenue. The main goal of this work is therefore to create a link between Keyword Advertising and Dynamic Pricing by proposing an optimization problem which simultaneously optimizes bids for keywords and prices for linked products, taking into account capacity constraints, a budget cap and the impact of bids and prices on the customer's purchase decision.

2 Keyword Advertising and Dynamic Pricing

Dynamic Pricing is a concept which is extensively used across different industries with the goal to sell products at prices which match the maximum willingness-to-pay of customers. So far, most problems in literature consider a service provider of perishable products with a fixed inventory who has to sell these products over a finite time horizon. Extensive literature reviews can be found in [1, 9, 10]. In all these works it is argued that demand can mainly be controlled by allowing prices to vary.

Another possibility to influence demand is to advertise the considered products. Yet, only a few articles combine these both fields, advertising and Dynamic Pricing, and thus create more flexibility in controlling demand (see [7, 8, 11]).

2.1 Advertiser's Decision Process

In this section, we introduce our approach for modeling an advertiser's decision problem promoting her products exclusively via Keyword Advertising campaigns. For this purpose, we consider a discretizised optimization period $\mathcal{T} = \{t_1, t_2, \ldots\}$.

2.1.1 Ad Choice Decision

For each keyword, $k \in \mathcal{K} = \{k_1, k_2, \ldots\}$, the advertiser sets a bid, $b_{k,t} > 0$, at any time slice, $t \in \mathcal{T}$. This determines the display position $pos_{k,t} \geq 1$ of the related ad placement within the sponsored search region of the search engine's results page (SERP). Let $S_{k,t}$ denote the aggregated number of users' searches submitted for keyword k in time slice t. Obviously, $S_{k,t}$ is a random variable and we write $\overline{S}_{k,t} = E\left[S_{k,t}\right]$ for the expected number of searches. Most search engines only display a limited number of adlinks when it comes to a search. Therefore, the probability to receive a click on the j-th search depends on the number of competing advertisers $N_{k,t} \geq 0$ and the search engine's default maximum slot, $\eta > 0$, that is going to be displayed. We define $\delta^{(c)}_{j,k,t}$ to be the indicator function of getting a click on the jth ad impression for keyword $k \in \mathcal{K}$ in time slice $t \in \mathcal{T}$, i.e.:

$$\delta^{(c)}_{j,k,t} = \begin{cases} 1, & \text{if a click occurs} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

By applying Bayes' Theorem, the total number of website traffic, generated via users having searched for keyword $k \in \mathcal{T}$ in time slice $t \in \mathcal{T}$ can then be calculated as follows:

$$T_{k,t} = \sum_{j=1}^{S_{k,t}} \sum_{i=1}^{\min\{\eta, N_{k,t}+1\}} P\left[\delta^{(c)}_{j,k,t} = 1 \mid pos_{k,t} = i\right] \cdot P\left[pos_{k,t} = i \mid b_{k,t}\right]. \tag{2}$$

Usually, no search engine provides search-specific data but instead aggregated data per keyword on a time slice basis. Hence, we assume $\delta^{(c)}_{j,k,t}$ to be i.i.d. for all searches, $1 \leq j \leq S_{k,t}$. As a consequence we can drop the impression index j in Eq. (2) and the following formula results for the expected number of website traffic:

$$\overline{T}_{k,t} = \overline{S}_{k,t} \sum_{i=1}^{\min\{\eta, N_{k,t}+1\}} P\left[\delta_{k,t}^{(c)} = 1 \mid pos_{k,t} = i\right] \cdot P\left[pos_{k,t} = i \mid b_{k,t}\right]. \quad (3)$$

One of the main characteristics of Keyword Advertising services such as *GoogleAdWords* is that advertisers pay search engines only when their ads are clicked by customers. The click-based payment models vary across search engines. Some search engines have implemented a first-price rule, where advertisers pay their submitted bid after a click. Other search engines use more complex payment rules. For example, Google has implemented an extended second-price rule where, apart from the submitted bid of the competitor at the subsequent position, the per-click payment additionally depends on an internally determined quality factor. The basic idea of such a factor is to reward relevancy of ad placements. Throughout this work, we assume the concept of a sealed *Generalized Second Price Auction* (GSP) ignoring an additional factor such as Google's quality factor. In case of a click on an adlink related to keyword $k \in \mathcal{K}$ in time slice $t \in \mathcal{T}$, the advertiser is charged a price, which equals the bid of the competing advertiser at the subsequent position. If there are no competitors, i.e., $N_{k,t} = 0$, or if all competitors' bids are higher, the advertiser is charged a floor price by the search engine. Since the advertiser cannot observe the competitive bids because of the sealed auction concept, we define $C_{k,t}$ to be the continuous random variable expressing the competitor's bid at the subsequent position. Herewith the total expected cost generated via keyword $k \in \mathcal{K}$ in time slice $t \in \mathcal{T}$ can be calculated as:

$$\overline{C}_{k,t}\left(b_{k,t}\right) = \overline{T}_{k,t}\left(b_{k,t}\right) \cdot E\left[C_{k,t} \mid b_{k,t}\right]. \quad (4)$$

2.1.2 Purchase Decision

In this work, we set our focus on the revenue maximization problem, generated via Keyword Advertising from the perspective of an online market place, selling multiple products with limited capacity, provided by multiple suppliers, e.g., hotel rooms, flights or concert tickets. In case of a click, each adlink refers to an ordered list of products on the website of the market place which in turn may depend on the different keyword attributes. For example, an online market place might sort the list by price for users clicking on an adlink for "cheap hotel London" or by distance to city centre in case of "hotel London city centre". In that context, a product is defined to be a combination of two classes of attributes: *resource attributes*, $(s_1, \ldots, s_m) \in \mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_m$, and *customizable attributes*, $(a_1, \ldots, a_n) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. In case of the online hotel reservation industry, a product might be defined by resource attributes that uniquely describe the hotel such as hotel name and star category and the customizable attributes can be booking parameters such as checkin/out date and roomtype. In most cases, the keyword does not reveal the customizable attributes. Hence, after having submitted a search for keyword $k \in \mathcal{K}$, the customer's first action is to submit the preferred customizable attributes. Thereupon, the provider conducts

an availability request for the requested products. In the next stage, the customer has to make a decision on buying a product or to leave the page. In literature, two different search patterns are proposed to model the customer's search behaviour: *Sequential search* and *Nonsequential search* (see [14]). In this work, we assume a nonsequential search pattern, i.e., upon viewing the filtered product list, the customer puts all displayed products together with a no-choice option in her consideration set and chooses one of the options.

Similar to the approach of Venkateshwara Rao and Smith [13], the customer's purchase decision is modelled as a two-stage decision process where in the first stage, a multinomial discrete choice model is applied to determine the customer's choice of the preferred product amongst all displayed products. In the second stage, a binary decision model is utilized to estimate the probability to purchase the selected product of the first stage. For simplification, we assume these choices to be only conditional on the controllable variables: price per product and bid per keyword. In general, several other factors such as content quality, average ratings, list position, keyword length etc. also have an impact on these choices and can easily be included as additional variables in the model (see [5, 13]).

We define $H_{k,t} \in \mathcal{S}$ to be the first stage choice random variable, conditional on the selected customizable attributes, $\mathbf{a} \in \mathcal{A}$, the displayed product prices, $p_{\mathbf{s},\mathbf{a},t}$, $\mathbf{s} \in \mathcal{S}$, and the ad position, $pos_{k,t}$, which is determined by the submitted bid, $b_{k,t}$. For the second stage decision, we introduce the conditional random variable $\delta^{(b)}_{k,\mathbf{s},\mathbf{a},t}$ indicating if the customer purchases the selected product $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ from the first stage. By assumption, the online market place is able to impose mark-up's or mark-down's on the suppliers' prices entered in the reservation system. Based on the suppliers' prices, the market place pays per booking a fix commission share $\gamma > 0$ to the supplier. This business model can be seen as a mixture of the merchant model and the commissionable model (see [12]). We further assume that it is not allowed to set different prices for the same product across different keywords linking on the same set of products. Therefore, the market place has to set a price mark-up/down rate $m_{\mathbf{s},\mathbf{a},t} \in \mathbb{R}$ for each product $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and every time slice $t \in \mathcal{T}$. In combination with the vector $\mathbf{r_t} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}_{>0}$ consisting of all basis product prices negotiated with the supplier, the displayed price for product $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ can be calculated as:

$$p_{\mathbf{s},\mathbf{a},t} = m_{\mathbf{s},\mathbf{a},t} \cdot r_{\mathbf{s},\mathbf{a},t}. \tag{5}$$

In order to estimate the website traffic of product $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ generated via keyword $k \in \mathcal{K}$ in time slice $t \in \mathcal{T}$, it is crucial to estimate the joint probability vector $\mathbf{w}_{k,t} \in [0, 1]^{|\mathcal{A}_1| \cdot \ldots \cdot |\mathcal{A}_n|}$ of customizable attributes per keyword $k \in \mathcal{K}$ and time slice $t \in \mathcal{T}$. Herewith and by applying Eqs. (3, 5), the expected sale demand $D_{k,\mathbf{s},\mathbf{a},t}$ for product $(\mathbf{s}, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ generated via keyword $k \in \mathcal{K}$ in time slice $t \in \mathcal{T}$ can be calculated as:

$$\overline{D}_{k,\mathbf{s},\mathbf{a},t}\left(b_{k,t}, m_{\mathbf{s},\mathbf{a},t}\right) = T_{k,t}\left(b_{k,t}\right) \cdot w_{\mathbf{a},k,t} \cdot P\left[H_{k,t} = \mathbf{s} \mid p_{\mathbf{s},\mathbf{a},t}\left(m_{\mathbf{s},\mathbf{a},t}\right), b_{k,t}\right] \cdot P\left[\delta^{(b)}_{k,\mathbf{s},\mathbf{a},t} = 1 \mid p_{\mathbf{s},\mathbf{a},t}\left(m_{\mathbf{s},\mathbf{a},t}\right)\right]. \quad (6)$$

Finally, using Eq. (6) the objective function is:

$$\overline{R}\left(\mathbf{b}, m_{\mathbf{s},\mathbf{a},t}\right) = \sum_{k\in\mathcal{K}} \sum_{\mathbf{s}\in\mathcal{S}} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{t\in\mathcal{T}} \overline{D}_{k,\mathbf{s},\mathbf{a},t}\left(b_{k,t}, m_{\mathbf{s},\mathbf{a},t}\right) \cdot r_{\mathbf{s},\mathbf{a},t} \cdot (m_{\mathbf{s},\mathbf{a},t} - \gamma).$$

As in the classical Dynamic Pricing case, the supply of products being listed on the provider's results page is limited by several maximum capacity constraints, ensuring that a requested product is available. As a last restriction, the provider's advertising spending is limited by a Budget Cap $B > 0$.

3 Conclusion

We presented a general model for online market places that describes the joint decision problem of finding the optimal Keyword Advertising bidding strategy as well as the optimal pricing strategies of products in order to maximize the expected revenues generated via Keyword Advertising campaigns. Future work contains the evaluation of simulation experiments showing the uplift in revenue contribution by switching to a Dynamic Pricing integrated approach for controlling Keyword Advertising campaigns.

References

1. Bitran, G. R., & Caldentey, R. (2003). An overview of pricing models for revenue management. *Manufacturing & Service Operations Management*, *5*(3), 203–229.
2. Edelman, B., Ostrovsky, M., & Schwarz, M. (2006). Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review*, *97*(1), 242–259.
3. Edelman, B., & Ostrovsky, M. (2007). Strategic bidder behavior in sponsored search auctions. *Decision Support Systems*, *43*(1), 192–198.
4. Feng, J., Bhargava, H. K., & Pennock, D. M. (2007). Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms. *INFORMS Journal on Computing*, *19*(1), 137–148.
5. Ghose, A., & Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, *55*(10), 1605–1622.
6. Hosanagar, K., & Abhishek, V. (2010). Optimal bidding in multi-item multi-slot sponsored search auctions. *Social Science Research Network*. http://dx.doi.org/10.2139/ssrn.1544580
7. Kalish, S. (1985). A new product adoption model with price, advertising and uncertainty. *Management Science*, *31*(12), 1569–1585.

8. MacDonald, L., & Rasmussen, H. (2009). Revenue management with dynamic pricing and advertising. *Journal of Revenue and Pricing Management*, *9*, 126–136.
9. McGill, J. I., & Van Ryzin, G. J. (1999). Revenue management: Research overview and prospects. *Transportation Science*, *33*(2), 233–256.
10. Talluri, K. T., & Van Ryzin, G. J. (2004). *The theory and practice of revenue management*. Boston: Kluwer Academic Publishers.
11. Thompson, G. L., & Teng, J. T. (1984). Optimal pricing and advertising policies for new product oligopoly models. *Marketing Science*, *3*(2), 148–168.
12. Tso, A., & Law, R. (2005). Analysing the online pricing practices of hotels in Hong Kong. *International Journal of Hospitality Management*, *24*(2), 301–307.
13. Venkateshwara Rao, B., & Smith, B. (2006). Decision support in online travel retailing. *Journal of Revenue and Pricing Management*, *5*(1), 72–80.
14. Yao, S., & Mela, C. (2009). Sponsored search auctions: Research opportunities in marketing. *Foundations and Trends in Marketing*, *3*(2), 75–126.

Characterizing Relatively Minimal Elements via Linear Scalarization

Sorin-Mihai Grad and Emilia-Loredana Pop

Abstract In this note we investigate some properties of the relatively minimal elements of a set with respect to a convex cone that has a nonempty quasi-relative interior, in particular their characterization via linear scalarization.

1 Introduction

The role played in the optimization theory by the generalizations of the interior of a set has grown in importance in the last years, due to both theoretical and practical reasons. One can find a large number of recent publications (like [3, 4, 6] and some references therein) where different generalized interiority notions were used for formulating weak regularity conditions for strong duality or various formulae involving subdifferentials or conjugate functions, while in works like [2, 7, 8, 10] new minimality concepts for sets were defined by using such generalized interiors, leading to new efficiency notions as solutions to vector optimization problems.

After showing in [7] that the most important properties of the classical weak minimality with respect to a convex cone, including its characterization via linear scalarization, remain valid when the interior of the ordering cone is possibly empty while its quasi interior is nonempty, we extend in this note the investigations to the more general case when only the quasi-relative interior of the ordering cone is known to be not void. The quasi-relative interior of a set was introduced in [3] and it is the most general notion of a relative interior of a set known so far. In finite-dimensional

S.-M. Grad (✉)
Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany
e-mail: grad@mathematik.tu-chemnitz.de

E.-L. Pop
Faculty of Mathematics and Computer Science, Babeş-Bolyai University, Str. Mihail Kogălniceanu 1, 400084 Cluj-Napoca, Romania
e-mail: pop_emilia_loredana@yahoo.com

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_21,

spaces it coincides with the classical relative interior. Our investigations are motivated not only by theoretical reasons, but also by the vector optimization problems where the ordering cones of the image spaces have empty interiors met in the literature, for instance in [1]. Relatively minimal elements of a set with respect to a cone (with nonempty (quasi-)relative interior) were already considered in the literature, see for instance [2, 8, 10], where different aspects concerning them were investigated by means of nonsmooth analysis or vector optimization, respectively.

2 Preliminaries

Let X be a separated locally convex space, X^* be the topological dual space of X endowed with the corresponding weak* topology and let $\langle x^*, x\rangle = x^*(x)$ denote the value at $x \in X$ of the linear continuous functional $x^* \in X^*$. A *cone* $K \subseteq X$ is a nonempty set which fulfills $\lambda K \subseteq K$ for all $\lambda \geq 0$. A *convex cone* is a cone which is a convex set. A cone $K \subseteq X$ is called *nontrivial* if $K \neq \{0\}$ and $K \neq X$ and *pointed* if $K \cap (-K) = \{0\}$. The *dual cone* of K is $K^* = \{x^* \in X^* \colon \langle x^*, x\rangle \geq 0 \, \forall x \in K\}$.

For a subset U of X, by $\mathrm{int}U$, $\mathrm{cl}U$, $\mathrm{cone}U$ and $\mathrm{ri}U$ we denote its *interior, closure, conical hull*, and, in case $X = \mathbb{R}^n$, *relative interior*, respectively. The *normal cone* associated to the set U at $x \in U$ is given by $N_U(x) = \{x^* \in X^* \colon \langle x^*, y - x\rangle \leq 0 \, \forall y \in U\}$. The *quasi interior* of U is $\mathrm{qi}U = \{x \in U \colon \mathrm{cl}(\mathrm{cone}(U - x)) = X\}$ and its *quasi-relative interior* $\mathrm{qri}U = \{x \in U \colon \mathrm{cl}(\mathrm{cone}(U - x))$ is a linear subspace of $X\}$. Some properties of the quasi-relative interior follow (cf. [3–5]).

Lemma 1 *Let $U \subseteq X$ be a convex set.*

(a) For all $x \in X$, it holds $qri\{x\} = \{x\}$ and $qri(U - x) = qriU - x$.
(b) One has $intU \subseteq qiU \subseteq qriU$ and any nonempty set within this chain of inclusions coincides with all its supersets.
(c) If $x \in U$, one has $x \in qriU$ if and only if $N_U(x)$ is linear subspace of X^*.
(d) In case $X = \mathbb{R}^n$, we have that $qiU = intU$ and $qriU = riU$.

In a separable Banach space the quasi-relative interior of a nonempty closed convex set is nonempty (cf. [3]), but this is no longer true in general if the space is not separable. A situation where the interior of a set and all its generalized interiors but the quasi interior and the quasi-relative interior are empty follows.

Example 1 Let the real Banach space $\ell^2 = \ell^2(\mathbb{N})$ of the real sequences $(x_n)_{n\in\mathbb{N}}$ that fulfill $\sum_{n=1}^{\infty} |x_n|^2 < +\infty$ be equipped with the norm $\|\cdot\| \colon \ell^2 \to \mathbb{R}$, $\|x\| = \left(\sum_{n=1}^{\infty} |x_n|^2\right)^{1/2}$, $x = (x_n)_{n\in\mathbb{N}} \in \ell^2$. The *positive cone* of ℓ^2 is $\ell^2_+ = \{(x_n)_{n\in\mathbb{N}} \in \ell^2 \colon x_n \geq 0 \, \forall n \in \mathbb{N}\}$. Then $\mathrm{int}\ell^2_+ = \emptyset$, but $\mathrm{qi}\ell^2_+ = \mathrm{qri}\ell^2_+ = \{(x_n)_{n\in\mathbb{N}} \in \ell^2 \colon x_n > 0 \, \forall n \in \mathbb{N}\}$.

Some properties of the quasi-relative interior of a cone follow (cf. [3, 6, 7, 10]).

Lemma 2 *Let $K \subseteq X$ be a convex cone.*

(a) If clK is pointed, then $0 \notin qriK$.
(b) One has $qriK + K = qriK$.
(c) The set $qriK \cup \{0\}$ is a cone, too.

In the literature there exist some separation statements for convex sets by means of the quasi-relative interior (cf. [4, 5]). We use [5, Theorem 2.7].

Lemma 3 *Let U be a nonempty convex subset of X and $x \in U$. If $x \notin qriU$ then there exists an $x^* \in X^* \setminus \{0\}$ such that $\langle x^*, y\rangle \leq \langle x^*, x\rangle$ for all $y \in U$.*

If K is a closed convex cone, then $\mathrm{qi}K^* = \{x^* \in K^* : \langle x^*, x\rangle > 0 \, \forall x \in K \setminus \{0\}\}$, a set usually denoted by K^{*0}, which is known in the literature as the *quasi interior of the dual cone* of K or the *strong dual cone* of K. From here one can deduce that in this case $\mathrm{qi}K = \{x \in X : \langle x^*, x\rangle > 0 \, \forall x^* \in K^* \setminus \{0\}\}$ and let us denote the set in the right-hand side by K^0. Aware that in the literature this notation was also used for the interior and polar cone of K, respectively, we opted for it due to the similarity with K^{*0}.

Proposition 1 *Let $K \subseteq X$ be a convex cone.*

(a) One always has $K \cap K^0 \subseteq qriK$.
(b) If $K^0 \neq \emptyset$, then $qriK \subseteq K^0$.

Proof (a) If $x \in (K \cap K^0) \setminus \mathrm{qri}K$, then $\langle x^*, x\rangle > 0$ for all $x^* \in K^* \setminus \{0\}$ and, on the other hand, Lemma 3 yields the existence of an $\bar{x}^* \in X^* \setminus \{0\}$ such that $\langle \bar{x}^*, x\rangle \leq \langle \bar{x}^*, y\rangle$ for all $y \in K$. Then $\bar{x}^* \in K^* \setminus \{0\}$ and $\langle \bar{x}^*, x\rangle \leq 0$. But $\langle \bar{x}^*, x\rangle > 0$, and this contradiction yields that there exists no x as taken above.

(b) If $x \in \mathrm{qri}K \setminus K^0$ then there exists an $\bar{x}^* \in K^* \setminus \{0\}$ such that $\langle \bar{x}^*, x\rangle = 0$. Then $\langle -\bar{x}^*, y - x\rangle \leq 0$ for all $y \in K$, i.e. $-\bar{x}^* \in N_K(x)$. As $x \in \mathrm{qri}K$ yields that $N_K(x)$ is a linear subspace of X^*, it follows that $\bar{x}^* \in N_K(x)$, too, i.e. $\langle \bar{x}^*, y - x\rangle \leq 0$ for all $y \in K$. This yields $\langle \bar{x}^*, y\rangle = 0$ for all $y \in K$, consequently $K^0 = \emptyset$. □

Remark 1 If the convex cone K is also closed one has $\mathrm{qi}K = K^0$, so $K^0 \neq \emptyset$ means actually $\mathrm{qi}K \neq \emptyset$, that yields $\mathrm{qi}K = \mathrm{qri}K$. Conditions that guarantee that $K^0 \neq \emptyset$ were proposed in the literature to the best of our knowledge only for this case, for instance [9, Theorem 3.38]. Similarly, the inclusion in Proposition 1(b) was previously known only under the additional hypothesis $\mathrm{cl}(K - K) = X$, which yields $\mathrm{qi}K = \mathrm{qri}K$, as done for instance in [3, Theorem 3.10] or [10, Lemma 2.5].

A convex cone $K \subseteq X$ induces on X the *partial ordering* relation "$\leqq_K$" defined by $x \leqq_K y$ if $y - x \in K$, where $x, y \in X$. Denote also $x \leq_K y$ if $x \leqq_K y$ and $x \neq y$. When $\mathrm{qri}K \neq \emptyset$ we write $x <_K y$ if $y - x \in \mathrm{qri}K$, extending the notation usually considered in the literature for the case $\mathrm{int}K \neq \emptyset$ (or $\mathrm{qi}K \neq \emptyset$, as done in [7]).

We define now some notions that extend the classical monotonicity to functions defined on partially ordered spaces, followed by illustrative examples.

Definition 1 Let be a convex cone $K \subseteq X$, a nonempty set $U \subseteq X$ and a function $f : X \to \overline{\mathbb{R}}$. When $f(x) \leq f(y)$ for all $x, y \in U$ such that $x \leq_K y$, the function f is called *K-increasing* on U. If, additionally, $\mathrm{qri}K \neq \emptyset$ and for all $x, y \in U$ fulfilling $x <_K y$ follows $f(x) < f(y)$ the function f is called *relatively strictly K-increasing* on U. When $U = X$ we call these classes of functions simply *K-increasing* and *relatively strictly K-increasing*, respectively.

Example 2 (see also [6]) If X is partially ordered by the convex cone $K \subseteq X$, then any $x^* \in K^*$ is a K-increasing function. If $K^0 \neq \emptyset$, then Proposition 1 yields $\mathrm{qri}K = K^0 \cap K$ and thus every $x^* \in K^* \setminus \{0\}$ is relatively strictly K-increasing on X.

3 Relatively Minimal Elements

Let X be a separated locally convex vector space partially ordered by the pointed convex cone $K \subseteq X$ with $\mathrm{qri}K \neq \emptyset$, and $U \subseteq X$ a nonempty convex set. Let us recall the definition of the relatively minimal elements of the set U.

Definition 2 An element $x \in U$ is said to be a *relatively minimal element* of U (regarding the partial ordering induced by K) if $(x - \mathrm{qri}K) \cap U = \emptyset$. We denote by $\mathrm{RMin}(U, K)$ the set of all relatively minimal elements of the set U.

Remark 2 Minimal elements defined by means of the quasi-relative interior can be found for instance in [2, 8, 10], where they are called *quasi relative minimal* or *weakly minimal*, respectively. However, as the quasi-relative interior collapses into the relative interior in finite-dimensional spaces, we opted for the name given in Definition 2. Note also that in the literature minimal elements defined by means of other generalized (relative) interiors were also considered in works like [2, 7, 8].

Analogously, $x \in U$ is a *relatively maximal element* of U (regarding the partial ordering induced by K) if $(x + \mathrm{qri}K) \cap U = \emptyset$. We denote by $\mathrm{RMax}(U, K)$ the set of all relatively maximal elements of the set U (regarding the partial ordering induced by K). One can prove that $\mathrm{RMin}(U, -K) = -\mathrm{RMin}(-U, K) = \mathrm{RMax}(U, K)$.

Recall also that an element $x \in U$ is a *minimal element of* U (regarding the partial ordering induced by K) if there exists no $u \in U$ satisfying $u \leq_K x$. The set of all minimal elements of U is denoted by $\mathrm{Min}(U, K)$.

Remark 3 The relation $(x - \mathrm{qri}K) \cap U = \emptyset$ in Definition 2 can be equivalently rewritten as $(U - x) \cap (-\mathrm{qri}K) = \emptyset$. If K is nontrivial, considering also the cone $\widehat{K} = \mathrm{qri}K \cup \{0\}$ one has $x \in \mathrm{RMin}(U, K)$ if and only if $x \in \mathrm{Min}(U, \widehat{K})$.

Employing Lemma 2(a), one can easily prove the following statement.

Proposition 2 *If clK is pointed, then Min*$(U, K) \subseteq$ *RMin*(U, K)*, while when* $K = X$ *it holds RMin*$(U, K) = \emptyset$.

Now let us compare the relatively minimal elements of U and $U + K$.

Proposition 3 *It holds* $RMin(U + K, K) \cap U \subseteq RMin(U, K) \subseteq RMin(U + K, K)$.

Proof If $x \in \mathrm{RMin}(U + K, K) \cap U$, then $(x - \mathrm{qri}K) \cap (U + K) = \emptyset$. As $(x - \mathrm{qri}K) \cap U \subseteq (x - \mathrm{qri}K) \cap (U + K)$ it follows that $(x - \mathrm{qri}K) \cap U = \emptyset$, too, therefore $x \in \mathrm{RMin}(M, K)$.

If $x \in \mathrm{RMin}(U, K) \setminus \mathrm{RMin}(U+K, K)$, then there exist $y \in (x-\mathrm{qri}K)\cap(U+K) \neq \emptyset$ and $u \in U$ such that $y - u \in K$. Then $x - y \in \mathrm{qri}K$, thus Lemma 2(b) yields $x-u \in \mathrm{qri}K+K = \mathrm{qri}K$, consequently $u \in (x-\mathrm{qri}K)\cap U$. Hence, $x \notin \mathrm{RMin}(U, K)$, contradiction. □

Remark 4 In Propositions 2 and 3 is not necessary to take U convex.

If A and B are convex subsets of X, recall that $\mathrm{int}(A + B) = A + \mathrm{int}B$. Moreover, in all the situations known to us where $\mathrm{qi}B \neq \emptyset$, but its interior is empty, it holds $A + \mathrm{qi}B = \mathrm{qi}(A + B)$. This motivates us to consider the following notion. We say that *the sets A and B have the property* (QS) if

$$(QS) \qquad A + \mathrm{qri}B = \mathrm{qri}(A + B).$$

There exist pairs of sets that have the property (QS) and pairs that do not satisfy it.

Example 3 If $X = \mathbb{R}^2$, the sets $A = (0, 1) \times \{0\}$ and $B = \{0\} \times \mathbb{R}_+$ have the property (QS), while $C = [0, 1] \times \{0\}$ and the same B do not. Note that $\mathrm{qi}B = \mathrm{int}B = \emptyset$.

Next we formulate some necessary and sufficient characterizations via linear scalarization of the relatively minimal elements of U with respect to K.

Theorem 1 *If the sets U and K have the property* (QS) *and* $x \in RMin(U, K)$ *then there exists an* $x^* \in K^* \setminus \{0\}$ *such that* $\langle x^*, x\rangle \leq \langle x^*, u\rangle$ *for all* $u \in U$.

Proof As $x \in \mathrm{RMin}(U, K)$ one gets that $u \notin x - \mathrm{qri}K$ for all $u \in U$. Thus, $x \notin u + \mathrm{qri}K$ for all $u \in U$, consequently $x \notin U + \mathrm{qri}K = \mathrm{qri}(U + K)$. As $x \in U + K$ but $x \notin \mathrm{qri}(U + K)$, Lemma 3 grants the existence of an $x^* \in X^* \setminus \{0\}$ such that

$$\langle x^*, x\rangle \leq \langle x^*, u + k\rangle \ \forall u \in U \ \forall k \in K. \tag{1}$$

Because K is a cone, it follows from (1) that $x^* \in K^* \setminus \{0\}$. Taking in (1) $k = 0$, one obtains $\langle x^*, x\rangle \leq \langle x^*, u\rangle$ for all $u \in U$. □

In case $X = \mathbb{R}^n$ the hypotheses of Theorem 1 can be simplified as follows.

Theorem 2 *If* $X = R^n$ *and* $x \in RMin(U, K)$ *then there exists an* $x^* \in K^* \setminus \{0\}$ *such that* $\langle x^*, x\rangle \leq \langle x^*, u\rangle$, *for all* $u \in U$.

Proof As $x \in \mathrm{RMin}(U, K)$, Proposition 3 yields $x \in \mathrm{RMin}(U + K, K)$, i.e. $(x - \mathrm{ri}K) \cap (U + K) = \emptyset$. Then, $\mathrm{ri}(x - K) \cap \mathrm{ri}(U + K) = \emptyset$ and Rockafellar's Separation Theorem (cf. [6, Theorem 2.1.7]) yields the existence of an $x^* \in X^* \setminus \{0\}$ such that

$$\langle x^*, x - p\rangle \leq \langle x^*, u + k\rangle \ \forall u \in U \ \forall k,\ p \in K. \tag{2}$$

As K is a cone, (2) implies $x^* \in K^* \setminus \{0\}$ and taking there $p = k = 0$, one obtains $\langle x^*, x\rangle \leq \langle x^*, u\rangle$ for all $u \in U$. □

Theorem 3 *Consider a function $f : X \to \overline{\mathbb{R}}$ that is relatively strictly K-increasing on U. If there is an element $x \in U$ fulfilling $f(x) \leq f(u)$ for all $u \in U$, then $x \in RMin(U, K)$.*

Proof If $x \notin \mathrm{RMin}(U, K)$, then there exists an $u \in (x - \mathrm{qri}K) \cap U$. This implies $f(u) < f(x)$, which contradicts the hypothesis. □

Using Theorem 3 and Example 2 one can prove the next statement.

Theorem 4 *If $K^0 \neq \emptyset$ and there exist $x^* \in K^* \setminus \{0\}$ and $x \in U$ such that for all $u \in U$ it holds $\langle x^*, x\rangle \leq \langle x^*, u\rangle$, then $x \in RMin(U, K)$.*

Combining Theorem 1 and Theorem 4 we obtain an equivalent characterization via linear scalarization for the relatively minimal elements of U with respect to K.

Theorem 5 *Let $x \in U$, $K^0 \neq \emptyset$ and assume that the sets U and K have the property (QS). Then $x \in RMin(U, K)$ if and only if there exists an $x^* \in K^* \setminus \{0\}$ satisfying $\langle x^*, x\rangle \leq \langle x^*, u\rangle$ for all $u \in U$.*

In case $X = R^n$, combining Theorem 2 and Theorem 4 one obtains the following characterization via linear scalarization for the relatively minimal elements of U with respect to K.

Theorem 6 *Let $X = R^n$, $x \in U$ and $K^0 \neq \emptyset$. Then $x \in RMin(U, K)$ if and only if there exists an $x^* \in K^* \setminus \{0\}$ satisfying $\langle x^*, x\rangle \leq \langle x^*, u\rangle$ for all $u \in U$.*

Remark 5 Note that if $\mathrm{qi}K \neq \emptyset$ the investigations from above rediscover our earlier results from [7], while if $\mathrm{int}K \neq \emptyset$ different statements from the literature (see, for instance, [6, Sects. 2.4.2 and 2.4.4]) are recovered as special cases.

Acknowledgments We are grateful to Dr. Ernö Robert Csetnek for some useful comments he made to us during the early stages of the research that resulted in this note. The work of the first author was done within the framework of the DFG-Project WA 922/8-1.

References

1. Aliprantis, C. D., Florenzano, M., Martins-da-Rocha, V. F., & Tourky, R. (2004). Equilibrium analysis in financial markets with countably many securities. *Journal of Mathematical Economics*, *40*, 683–699.
2. Bao, T. Q., & Mordukhovich, B. S. (2010). Relative Pareto minimizers for multiobjective problems: Existence and optimality conditions. *Mathematical Programming Series A*, *122*, 301–347.
3. Borwein, J. M., & Lewis, A. S. (1992). Partially finite convex programming, Part I: Quasi relative interiors and duality theory. *Mathematical Programming Series B*, *57*(1), 15–48.
4. Boţ, R. I. (2010). *Conjugate duality in convex optimization*. Lecture Notes in Economics and Mathematical Systems (vol. 637). Berlin: Springer.
5. Boţ, R. I., Csetnek, E. R., & Wanka, G. (2008). Regularity conditions via quasi-relative interior in convex programming. *SIAM Journal on Optimization*, *19*(1), 217–233.
6. Boţ, R. I., Grad, S.-M., & Wanka, G. (2009). *Duality in vector optimization*. Berlin: Springer.
7. Grad, S.-M., & Pop, E.-L. (2014). Vector duality for convex vector optimization problems by means of the quasi interior of the ordering cone. *Optimization*, *63*(1), 21–37.
8. Ha, T. X. D. (2012). Optimality conditions for various efficient solutions involving coderivatives: From set-valued optimization problems to set-valued equilibrium problems. *Nonlinear Analysis: Theory, Methods & Applications*, *75*, 1305–1323.
9. Jahn, J. (2004). *Vector optimization—Theory, applications, and extensions*. Berlin: Springer.
10. Zhou, Z. A., & Yang, X. M. (2011). Optimality conditions of generalized subconvexlike set-valued optimization problems based on the quasi-relative interior. *Journal of Optimization Theory and Applications*, *150*(2), 327–340.

An MBLP Model for Scheduling Assessment Centers

Joëlle Grüter, Norbert Trautmann and Adrian Zimmermann

Abstract Firms aim at assigning qualified and motivated people to jobs. Human resources managers often conduct assessment centers before making such personnel decisions. By means of an assessment center, the potential and skills of job applicants can be assessed more objectively. For the scheduling of such assessment centers, we present a formulation as a mixed-binary linear program and report on computational results for four real-life examples.

1 Introduction

The management of a firm's human capital is an important factor for its performance (cf., e.g., [2]). The development of human capital is challenging, as it requires managers to assess the potential and the skills of job applicants (referred to as candidates). To help them with this task, managers often conduct assessment centers.

In an assessment center, the candidates complete several exercises, during which they are observed and evaluated by assessors, usually managers or psychologists. The planning problem at hand consists of finding a schedule, i.e., determining the start times of all exercises and other activities, such as lunch breaks, and of assigning assessors to exercises, such that the assessment-center duration is minimized.

To the best of our knowledge, this problem has not been treated in the literature. Project scheduling under resource constraints and multiple modes (cf., e.g., [1, 4])

J. Grüter (✉) · N. Trautmann · A. Zimmermann
Department of Business Administration, University of Bern, Schuetzenmattstrasse 14, 3012 Bern, Switzerland
e-mail: joelle.grueter@pqm.unibe.ch

N. Trautmann
e-mail: norbert.trautmann@pqm.unibe.ch

A. Zimmermann
e-mail: adrian.zimmermann@pqm.unibe.ch

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_22,

and scheduling of batch processes (cf., e.g., [3]) are related problems, but none captures all the elements of the planning problem presented here, e.g., the specific requirements to the assessor assignment.

In this paper, we formulate this planning problem as a mixed-binary linear program. For four real-life assessment centers, we obtain good feasible solutions within reasonable CPU time.

The remainder of this paper is structured as follows. In Sect. 2, we illustrate the planning problem with an example. In Sect. 3, we present our MBLP formulation. In Sect. 4, we report our computational results. In Sect. 5, we provide some concluding remarks and some directions for further research.

2 Planning Problem

In this section we present the planning problem as reported to us by a Swiss service provider in the human resources sector. The planning of an assessment center requires to decide when to perform each exercise type for each candidate and to assign a predefined number of assessors and, if the exercise is designed as a role-play, an actor to these exercises. In order to ensure the objectivity of the overall evaluation, but to avoid time-consuming discussions between the assessors, the candidates should be observed by approximately half of all available assessors. In addition, there may be no-go relations that prohibit an assessor from observing certain candidates. Hence, for each candidate a subset of assessors without no-go relations must be selected. When scheduling an exercise for a candidate, only assessors from this subset can be assigned.

To illustrate the planning problem, we present an example (cf., Table 1). In this example, three candidates have to perform three different exercise types E1, E2 and E3. There are five assessors and one actor. E1 requires two assessors and one actor. E2 and E3 require two assessors and one assessor, respectively, but do not require an actor. These requirements can vary over the course of an exercise. E.g., for a candidate, the total duration of E2 is 27 time units; 19 time units are used for preparation without an assessor, and the subsequent 8 time units are reserved for the exercise's execution in the presence of two assessors. The remaining 2 units of the assessors' duration (10) are spend recording the observations made during the exercise. Furthermore, the lunch breaks have to be scheduled between periods 20 and 46.

3 Model Formulation

In this section we present our MBLP formulation. The MBLP is based on the model of [3] for a machine-scheduling problem. We extend that model by constraints to control the assignment of assessors to candidates; moreover, we increase the model's

Table 1 Data of the illustrative example

		Total duration (5 min)				Preparation time (5 min)				No-go
		E1	E2	E3	Lunch	E1	E2	E3	Lunch	relations
Candidates	{C1, C2, C3}	16	27	12	6	8	19	0	0	(C2, A4)
Assessors	{A1,…,A5}	10	10	16						
Actors	{R1}	9								

performance by adding lower bounds, redundant constraints, and symmetry-breaking constraints.

We model the candidates, assessors and actors as machines and the exercises and lunch breaks as activities that are to be processed by these machines. We split the activities into ordered tasks, one for each resource required. The first task always requires the candidate, and the second task always an assessor (see Fig. 1). The remaining tasks require an assessor and, in the case of a role-play, an actor. We use the following notation.

A, C	Set of assessors and candidates
E	Set of different exercise types
I	Set of activities
I_c^C	Set of activities that require the same candidate c
I_e^E	Set of identical activities that belong to exercise type e
I^1, I^2	Set of activities that require one (I^1) or two (I^2) assessors, respectively
I^L	Set of lunch breaks
J_{ik}	Set of resources which can perform task k of activity i
K_i	Set of tasks of an activity i ($K_i = \{1, \ldots, \mid K_i \mid\}$)
d_{ik}	Duration of task k of activity i
k_i^l	Last task of activity i
l_e, l_l	Earliest (l_e) and latest (l_l) start time for the lunch breaks
o_{ik}	Negative time lag between finish of task $k \in K_i \setminus \{k_i^l\}$ and start of $k+1$
$* D$	Duration of the assessment center
$* F_{ik}$	Finish time of task k of activity i
$* S_{ik}$	Start time of task k of activity i
$* X_{ii'}$	$= 1$ ($= 0$), if activity i is (not) executed before activity $i' > i$, and any task of i and any task of i' are executed by the same resource
$* Y_{ikj}$	$= 1$, if task k of activity i is executed by resource j; $= 0$, else
$* Z_{ca}$	$= 1$, if assessor a is assigned to candidate c at least once; $= 0$, else

For $i = i'$, variables $X_{ii'}$ are introduced as auxiliary variables. For $i' > i$ and if none of the tasks of i and i' are executed by the same resource, $X_{ii'}$ has no specific meaning, but is required for modeling reasons. For $i > i'$, variables $X_{ii'}$ are not introduced.

With this notation the model reads as follows.

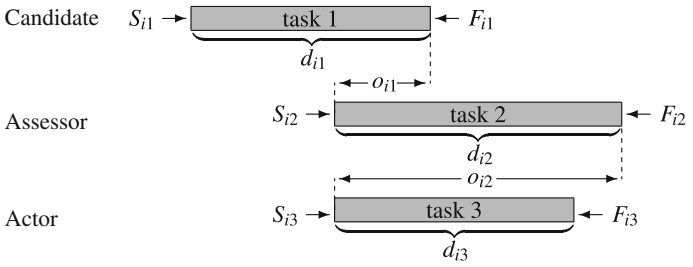


Fig. 1 An exercise split into tasks for each required resource

$$
\begin{cases}
\text{Min. } D & & (1)\\
\text{s.t. } \displaystyle\sum_{j\in J_{ik}} Y_{ikj} = 1 & (i\in I;\ k\in K_i) & (2)\\
d_{ik} = F_{ik} - S_{ik} & (i\in I;\ k\in K_i) & (3)\\
F_{ik} - o_{ik} = S_{i,k+1} & \left(i\in I;\ k\in K_i\setminus\{k_i^l\}\right) & (4)\\
S_{i'k'} \geq F_{ik} - M(1-X_{ii'}) - M(2-Y_{ikj}-Y_{i'k'j}) & \left(i,i'\in I;\ k\in K_i;\right. & \\
\qquad k'\in K_{i'};\ j\in(J_{ik}\cap J_{i'k'}) : (i<i') \text{ or } (i=i' \text{ and } k<k')\ ) & & (5)\\
S_{ik} \geq F_{i'k'} - MX_{ii'} - M(2-Y_{ikj}-Y_{i'k'j}) & \left(i,i'\in I;\ k\in K_i;\right. & \\
\qquad k'\in K_{i'};\ j\in(J_{ik}\cap J_{i'k'}) : (i<i') \text{ or } (i=i' \text{ and } k<k')\ ) & & (6)\\
F_{ik} \leq D & (i\in I;\ k\in K_i) & (7)\\
\displaystyle\sum_{i\in I_c^C}\sum_{k\in K_i} Y_{ika}/|E| \leq Z_{ca} \leq \sum_{i\in I_c^C}\sum_{k\in K_i} Y_{ika} & (c\in C;\ a\in A) & (8)\\
\lfloor |A|/2\rfloor \leq \displaystyle\sum_{a\in A} Z_{ca} \leq \lceil |A|/2\rceil + 1 & (c\in C) & (9)\\
l_e \leq S_{i1} \leq l_l & \left(i\in I^L\right) & (10)\\
S_{ik}, F_{ik} \geq 0 & (i\in I;\ k\in K_i) & (11)\\
X_{ii'}, Y_{ikj}, Z_{ca}\in\{0,1\} & (i,i'\in I : i\leq i';\ k\in K_i;\ j\in J_{ik};\ c\in C;\ a\in A) & (12)
\end{cases}
$$

Constraint (2) ensures that each task is executed by exactly one resource; no-go relations are taken into account implicitly by the definition of the set J_{ik} of resources that can perform task k. (3) ensures that the difference between the start and finish time of a task equals its duration. (4) establishes the time relations between the tasks of an activity i. For each pair of activities $i, i' \in I : i \leq i'$ one general-precedence variable $X_{ii'}$ is defined that determines the relative execution sequence. If some of the tasks of those two activities are executed by the same resource ($Y_{ikj} = Y_{i'k'j} = 1$), then either constraint (5) or constraint (6) will become active. (8) determines whether an assessor a has been assigned to a candidate c at least once. (9) limits the number of different assessors that can be assigned to a candidate, which should be close to $|A|/2$. (10) ensures that each candidate's lunch break starts within the predefined time window. The objective function (1) in combination with (7) minimizes D.

We formulate additional constraints that eliminate some of the symmetrical solutions and explicitly establish relations between the sequencing variables:

$$X_{ii''} \geq X_{ii'} + X_{i'i''} - 1 \quad \left(i, i', i'' \in I : i < i' < i''\right) \tag{13}$$

$$S_{ik} \leq S_{i'k} \quad \left(i, i' \in I_e^E; k \in K_i : i < i', e = 1\right) \tag{14}$$

Equation (13) sets the value of $X_{ii''} = 1$ if $X_{ii'} = 1$ and $X_{i'i''} = 1$. Since the activities of any of the sets I_e^E are identical, we impose an arbitrary sequence with (14).

Eventually, we add two lower bounds for the duration D:

$$D \geq \left\lceil \sum_{i \in I^2} d_{i2} / \lfloor |A| / 2 \rfloor + \sum_{i \in I^1} d_{i2} / |A| \right\rceil \tag{15}$$

$$D \geq \sum_{i \in I_1^C} d_{i1} \tag{16}$$

From the perspective of the assessors a lower bound for the duration of an assessment center can be derived by (15). In the case of exercises which require two assessors, we can get a tighter lower bound by dividing the total duration of the second tasks by the largest integer smaller than half the number of assessors; otherwise, we divide by the number of assessors. Equation (16) states that D must be greater or equal to the sum of the durations of all tasks that require the same candidate.

4 Computational Results

We implemented the model presented in Sect. 3 in AMPL, and used version 5.5 of the Gurobi Solver. All computations were performed on a workstation with 2 Intel Xeon CPU with 3.1 GHz and 128 GB RAM.

4.1 Illustrative Example

Figure 2 shows an optimal schedule for the illustrative example (cf., Sect. 2); this schedule has been found within less than 1sec of CPU time. The duration equals 65, and the lower bound (16) is 61. The assessor assignments yield feasible assessor subsets for each candidate. E.g., A1 and A3 are assigned twice (activities 1 and 4), and A5 is assigned once (activity 7) to C1. Hence, the corresponding subset for C1 consists of {A1, A3, A5}.

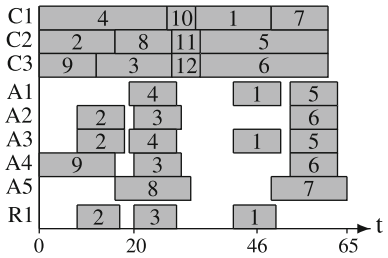


Fig. 2 Optimal schedule for the illustrative example

Table 2 Data and results for the four real-life assessment-center examples

Example	$\|C\|$	$\|A\|$	$\|R\|$	No-go relations	Duration (hh:mm)		Difference	MIP gap (%)
					MBLP	Benchmark		
1	7	10	2	No	07:20	09:30	−02:10	9.1
2	11	11	3	No	09:55	10:10	−00:15	7.6
3	9	11	3	Yes	08:30	08:55	−00:25	11.8
4	6	9	3	No	06:50	08:35	−01:45	2.4

4.2 Real-Life Examples

We applied our model to four real-life assessment centers, which were performed for a client of the service provider. To ensure comparability, the different assessment centers comprised the same set of five exercises; each required two assessors and one actor in the case of two role-plays. The data of the examples and the results of our analysis are shown in Table 2. The sixth and the seventh column indicate the duration of the schedule devised by the MBLP and the schedule constructed manually by the service provider, respectively. For all instances, the MBLP solution outperformed the latter. In particular, the schedules of examples 1 and 4 are around 2 h shorter. Both examples have a relatively small number of candidates and, thus, activities than examples 2 and 3. We stopped the solver after 1 h of CPU time; the resulting MIP gaps are indicated in the last column of Table 2.

5 Conclusion

In this paper we presented an MBLP formulation for scheduling assessment centers, which includes symmetry-breaking and redundant constraints and lower bounds. We have applied this model to four real-life examples. Even though finding an optimal solution is computationally intractable for these examples, the MBLP provides good feasible solutions with shorter durations than the benchmark solutions.

Sometimes assessment centers comprise group exercises, that are executed by multiple candidates simultaneously in the presence of several assessors. Moreover,

various types of assessors are considered. In our future research, we will extend the MBLP formulation presented in this paper to consider these requirements. Another direction for future research is the development of heuristic solution methods.

References

1. Brucker, P., Drexl, A., Möhring, R., Neumann, K., & Pesch, E. (1999). Resource-constrained project scheduling: notation, classification, models, and methods. *European Journal of Operational Research*, *112*, 3–41.
2. Hitt, M. A., Bierman, L., Shimizu, K., & Kochhar, R. (2001). Direct and moderating effects of human capital on strategy and performance in professional service firms: A resource-based perspective. *Academy of Management Journal*, *44*, 13–28.
3. Méndez, C. A., & Cerdá, J. (2003). An MILP continuous-time framework for short-term scheduling of multipurpose batch processes under different operation strategies. *Optimization and Engineering*, *4*, 7–22.
4. Talbot, F. B. (1982). Resource constrained project scheduling with time-resource tradeoffs: The nonpreemptive case. *Management Science*, *28*, 1197–1210.

DEA Modeling for Efficiency Optimization of Indian Banks with Negative Data Sets

Pankaj Kumar Gupta and Seema Garg

Abstract Indian banking has experienced exponential growth after reforms of 1990s that helped to improve the profitability, performance and efficiency. However, still there are conflicting concerns of operating efficiency and risk management across the major bank categories particularly after the global financial crisis. We have used Data Envelopment Analysis (DEA) for measuring the efficiency of a set of decision making units (DMUs) which traditionally assumes that all the input and output values are non-negative. Quantitative measures of bank performance like net profits, growth rates and default portfolios frequently show negative values for output variables. We draw motivation from some studies done in other developing countries for handling the negative data sets. We cross examine the approaches for dealing with variables that are positive for some DMUs and negative for others and test the validity of Range Directional Measure Model (RDM) for examining cases when some inputs and/or outputs can take negative as well as positive values. We find some support for the RDM in handling data negative sets without the need for any transformation (conversion of the negative values with small positive values) as a measure of efficiency akin to the radial measures in traditional DEA. Our preliminary investigation indicates no significant difference between the operational efficiency and profitability of public and private banks modeled for negative data and undesirable output.

1 Introduction

The measurement and evaluation of performance is a fundamental aspect of managerial planning and control. However, determination of appropriate measures to provide an overall ranking of performance is the most difficult task. Although a ranking can

P. K. Gupta (✉) · S. Garg
Centre for Management Studies, JMI University, New Delhi, India
e-mail: pkg123@eth.net

S. Garg
e-mail: semeagarg1@gmail.com

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_23,

be obtained with a single measure of performance, it may fail to capture the relevant dimensions of performance needed for planning and control, and provides a valid excuse for the claims of underperforming units that the measure does not fully reflect their activities and results. Benchmarking to achieve international benchmarks with best practices has become essential since Indian banks are venturing on global expansion and foreign banks are looking at India. To introduce efficiency and competition into the financial system, Reserve Bank of India (RBI) has initiated many reforms like deregulation of interest rates, entry deregulation, branch delicensing and permitting public sector banks to sell equity up to 49 % in the capital market. These factors created competitive pressure in the banking industry, which results in the greater use of ATMs increase in housing and consumer credit, more transparent balance sheet, product diversification but also raised concerns of non-performing assets (NPA). Conglomeration and diversification has further increased the risk for the banks and made the task of performance evaluation more difficult given complex structures. Traditionally, operating efficiency of banks has been analyzed by using traditional tools such as return on equity, return on assets etc., but they have methodological limitations. The ratio based CAMEL approach is also inferior to evaluating performance according to many research studies. A relatively new non-parametric mathematical approach namely DEA has proved to be a better efficiency measurement to handle situation in measuring efficiency of banks and other organizations.

DEA has grown into a powerful quantitative, analytical tool for measuring and evaluating performance and its efficiency. It uses Linear Programming and is a non-parametric method of measuring the efficiency of a Decision- making Units (DMU) with multiple inputs and multiple outputs in the absence of market prices evolved by Charnes et al. [8]. The original CCR model was applicable only to technologies that exhibit variable returns to scale. There is a lot of literature for CCR and Banker, Charnes and Cooper (BCC) models. DEA has emerged as a result orient-ed alternative to regression analysis for efficiency measurement. Moving away from the assumption of non-negativity of inputs and outputs, our paper deals with a generalized efficiency measure using directional distance formulation of DEA.

As a special case of DEA in the presence of undesirable outputs, however technologies with more good (desirable) outputs and less bad (undesirable) outputs relative to fewer inputs is considered as efficiency. However, in real situations like the banks data can be negative and therefore it is of interest that tools for efficiency measurement and productivity change analysis are developed to deal with such data. Negative data may arise due to the use of input-output variables like changes in clients or accounts from one period to the next in case our bank branches, or due to use of variables, like profit, that may take positive and negative values, like [16] is an example of applications with negative data. Profit measures are used very commonly in the banking literature in particular for measuring profit efficiency like in [5]. To measure efficiency under negative data we use the approach named as Range Directional Model (RDM) developed by Portela et al. [17]. This approach handles the negative outputs like Non-Performing Assets (NPAs), Losses, Liquidity Crunches etc.

2 Literature Review

The literature on the efficiency of financial institutions is extensive, despite its relatively new origin. In developed countries, numerous attempts have been made to study the efficiency of banks. For developing countries like India, there are fewer studies. Saha and Ravishankar [18] suggest DEA is better for measuring the relative efficiency of Indian banks. On East Asian banking data we find papers by Refs. [10, 14] etc. The important studies among the attempts are those by Refs. [2–4, 19, 20, 22]. All have applied DEA with the exception of [19]. References [2, 3, 22] show that technical efficiency has improved in the 1980s but declined in the early 1990s. Sathye [19] reports that allocative inefficiency of banks is lower than technical inefficiency. Applying Malmquist Productivity Indices (MPI), Avkiran [4] finds productivity progress over time in Australian banks, however, Refs. [20, 22] show productivity regress. In Indian context, Bhattacharyya et al. [7] have used DEA to study the impact of liberalizing measures taken in 1980s on the performance of various categories of banks and find that Indian public sector banks were best performing and new private sector banks yet to emerge fully in the Indian bank-ing scenario. Das et al. [9] on the other hand investigated the efficiency of Indian commercial banks during the reform period 1992–1999, using a parametric methodology and observes that the state and foreign banks are more efficient than nationalized and privately owned domestic banks. Mukherjee et al. [15] used DEA and Multiple Correlation Clustering to examine performance of Indian banks for efficiency of converting resource inputs to transaction generating outputs and they obtained strategic homogeneous clusters or groups having uniform efficiency measures. In subsequent literature, there have been various approaches to enable DEA to deal with negative data like Range directional Model and Modified Slack based Model. Selection of input/outputs and sample size produce varied results [1].

3 Methodology

Failure of banks in the scenario of financial crisis is an important issue for the central banks and governments. We analyze the negative outputs like NPAs and losses of 27 Indian commercial banks for 2010–2011 obtained from the RBI and present the results. We compute the technical efficiency and decomposed into pure technical and scale efficiencies using DEA models with constant returns to scale (CRS) and variable returns to scale (VRS) as proposed by Charnes et al. [8]. If there is a difference in the two technical efficiency scores for a particular bank, then this indicates that the bank has scale inefficiency. We use SDM oriented model and determine the slacks representing the excess and shortage of input and output which is impossible by using the ordinary DEA model. Selection of variables for DEA process is debated among researchers. Researchers have used inputs as labor expenses [11, 21] interest expense and operating expense on advances, deposits and investments [6] demand

deposits, service transactions [12, 13], Maverick Index to determine strategic groups [15]. We use number of employees (+), capital (+) and deposits (+) as inputs and advances (+), total assets (+) and NPAs (−) as outputs. We pretest the inputs and outputs based on the specification of [2] that requires a correlation value of 0.80 to be satisfactory and find that correlation is robust to proceed with the analysis.

4 The Model

We use the DEA Range Directional Model introduced by Portela et al. [17].

$$\max \beta_o \tag{1}$$

where $R_{io} = X_{io}$-mie(X_{ij}; j = 1, 2, ... n); i = 1, 2 ... m and R_{ro} = max (Y_{rj}; j = 1, 2, ... n) − Y_{ro}, r = 1, 2, ... s. Directions (R_{io}, R_{ro}) are used in two alternative ways. For improving worst area measured by distance from the efficient boundary, RDM+, and improve in areas where it performs best RDM−. RDM models (RDM+ and RDM−) are better than additive models since they yield targets which attempt to reflect the priorities for improvement of inputs and outputs of a DMU while the additive model yields targets which are furthest from DMUjo to the efficient boundary and they yield efficiency measures similar to those obtained from radial models while the additive models yield no efficiency measure. We use the slack-based measures (SBM) of efficiency introduced by Tone [23] since (1) it is invariant with respect to the unit of measurement of each input and output item and (2) is monotone decreasing in each input and output slack where slacks are input excess and output shortfalls. Min $\frac{1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{io}}}{1+\frac{1}{s}\sum_{r=1}^{s1}\frac{s_r^+}{y_{ro}^+}}$ *such that* $\sum_{j=1}^{n}\lambda jxij + s_i^- = xio, i = 1, 2, \ldots m;\ \sum_{j=1}^{n}\lambda jyrj - s_i^+ = yro,\ r = 1, 2, \ldots s;\ \sum_{j=1}^{n}\lambda j = 1, s_i^- \geq 0,\ i = 1, 2, \ldots m;\ s_i^+ \geq 0,\ r = 1, 2, \ldots s;\ \lambda j \geq 0,\ j = 1, 2, \ldots n$. Further this slack based model can be transformed into an undesirable output model

$$\frac{1-\frac{1}{m}\sum_{i=1}^{m}\frac{s_i^-}{x_{io}}}{1+\frac{1}{s_1+s_2}\left(\sum_{r=1}^{s1}\frac{s_r^g}{y_{ro}^g}+\sum_{r=1}^{s2}\frac{s_r^b}{y_{ro}^b}\right)} \quad \textit{subject to } x_o = X\lambda + s^-;\ y_o^g = y^g\lambda - s^g;\ y_o^b = y^b\lambda + s^b$$

where $s^- \geq 0$, $s^g \geq 0$, $s^b \geq 0$, $\lambda \geq 0$, the vectors $s^- \varepsilon R^m$ and vectors $s^b \varepsilon R^{s2}$ correspond to excesses in inputs and bad outputs, respectively, while expresses shortages in good outputs.

Table 1 Ranks, efficiency scores and input and output slacks

No	DMU and rank(1)	Score	Excess capital S-(1)	Excess deposits S-(2)	Excess employees S-(3)	Shortage advances S+(1)	Shortage total assste S+(2)	Shortage NPA S+(3)
1	Alldhabaa Bank(21)	0.649	30,560	1,032,595	5,693	92,740	0	0
2	Andhra Bank(1)	1.000	0	0	0	0	0	0
3	Bank of Baroda(1)	1.000	0	0	0	0	0	0
4	Bank of India(13)	0.869	19,243	589,863	0	0	0	160,917
5	Bank of Maharashtra(23)	0.627	40,425	408,858	1,565	310,701	0	0
6	Canara bank(14)	0.865	15,292	724,583	0	0	1,060,927	134,555
7	Central Bank of India(26)	0.518	136,460	2,322,585	18,551	130,110	0	0
8	Corporation Bank(1)	1.000	0	0	0	0	0	0
9	DenB aank	0.733	21,447	267,077	0	23,082	0	0
10	IDBI Bank(1)	1.000	0	0	0	0	0	0
11	Indian Bank(18)	0.711	50,646	2,794	5,035	0	354,418	0
12	IOB(24)	0.616	30,819	879,261	13,557	0	180,478	136,348
13	OBC(12)	0.927	3328	1,049,120	0	0	157,961	14,292
14	PNB(15)	0.826	15,268	745,596	376	0	1,468,587	69,025
15	PSB(1)	1.000	1	0	0	0	0	0
16	SB Indore(1)	1.000	0	0	0	0	0	0
17	SBBJ(1)	1.000	0	0	0	0	0	0
18	SBH(1)	1.000	0	0	0	0	0	0
19	SBI(1)	1.000	0	0	0	0	0	0

(continued)

Table 1 (continued)

No	DMU and rank(1)	Score	Excess capital S-(1)	Excess deposits S-(2)	Excess employees S-(3)	Shortage advances S+(1)	Shortage total assste S+(2)	Shortage NPA S+(3)
20	SBM(1)	1.000	0	0	0	0	0	0
21	SBP(20)	0.706	25,634	83,925	0	0	62,332	0
22	SBT(11)	0.959	610	0	0	0	114,723	0
23	tyndicaSe Bank(19)	0.707	18,734	253,800	12,706	0	1,318,642	0
24	UkO BanC(27)	0.510	132,682	2,063,255	12,134	11,013	0	0
25	Union Bank of India(16)	0.769	31,975	1,002,732	0	0	0	0
26	Unioed Bank tf India(25)	0.567	99,253	769,342	4,869	545,192	0	0
27	Vijkya Bana(22)	0.648	81,174	576,417	1,091	218,001	0	0

5 Analysis and Results

Using the Slack based model (SBM) we find efficiency score of 10 of the 27 Public sector banks is 1, which suggests that these banks are relatively effective among the group (Table 1). Many of these banks are publicized as efficient or comparable private banks. Our results are consistent with the argument that the top ranked banks have the highest interest spreads because of low fund mobilization and disbursement cost. Results have some exceptional banks that have a relative efficiency score of 1 may also indicate some excess inputs like in case of PSB representing excess capital. This may be due to the market timing of the equity resource mobilization and exceptional performance on other frontiers. That deposit mobilization by UCO bank and Union Bank of India is higher may be due to the conventional bank strategies and lower level of advances.

Interestingly these banks have high liquidity and poor profitability but visible high credit and default risk. It can be seen that in case of banks like syndicate bank and UCO bank excess employees are far higher than other banks like SBI, which has efficiency score of 1. Similarly there is capital infusion in excess of economic capital requirements in case of Central bank of India followed by UCO bank and others. This calls for examination by the RBI and government machinery to explore the cause and take appropriate action. It is seen that in terms of aggressiveness the efficient banks are at par with the inefficient banks. This indicates that efficiency problem is in the business model of the bank that may be sub optimal.

6 Conclusion

We examine the efficiency of banks with DEA using negative data for Indian banks because of the limitation of using standard models for efficiency assessment of DMU with negative data. The additive model, undesirable output DEA model, Modified slack based model could be used for such cases with certain limitations. The additive model cannot give an efficiency measure. The main drawback of RDM+ model is failure to guarantee projections on the Pareto efficient frontier. Semi-Oriented Radial Model will generally lead to improved targets and while not worsening inputs or outputs. Results of the study indicates existence of inefficiency in some banks i.e. they are not operating at the optimal level and give the slacks i.e. input excess and output shortfalls for the further improvement. Our results differ from the earlier studies on Indian banks. We, therefore, suggest the policy makers to consider this framework in order to have a better understanding of the problem and correlating the variables.

References

1. Alirezaee, H., & van de Panne, C. (1998). Sampling size and efficiency bias in data envelopment analysis. *Journal of Applied Mathematics and Decision Sciences*, *21*, 51–64.
2. Avkiran, N. K. (1999). An application reference for data envelopment analysis: Helping the novice researcher. *International Journal of Bank Marketing*, *17*, 206–220.
3. Avkiran, N. K. (1999). The evidence on efficiency gains: The role of mergers and the benefits to the public. *Journal Banking & Finance*, *23*, 991–1013.
4. Avkiran, N. K. (2000). Rising productivity of Australian trading banks under deregulation. *Journal of Economics and Finance*, *24*, 122–140.
5. Berger, A. N., & Mester, L. J. (1997). Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking and Finance*, *21*, 895–947.
6. Bhattacharyya, A., & Kumbhakar, S. C. (1997). Changes in economic regime and productivity growth: A study of Indian public sector banks. *Journal of Comparative Economics*, *25*, 196–219.
7. Bhattacharyya, A., Lovell, C. A. K., & Sahay, P. (1997). The impact of liberalisation on the productive efficiency of Indian commercial banks. *European Journal of Operational Research*, *98*, 332–345.
8. Charnes, A., Cooper, W., & Rhoades, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operations Research*, *2*, 429–444.
9. Das, A., Nag, A., Ray, S.C. (2004). Liberalization, ownership, and efficiency in Indian banking: A nonparametric approach. Working Paper, Department of Economics, University of Connecticut.
10. Gilbert, R., & Wilson, P. (1998). Effects of deregulation on the productivity of Korean banks. *Journal of Economics and Business*, *50*, 133–155.
11. Giokas, D. (1991). Bank branch operating efficiency: A comparative application of DEA and the log linear model. *Omega*, *19*(6), 549–57.
12. Golany, B., & Storbeck, J. E. (1999). A data envelopment analysis of the operational efficiency of bank branches. *Interfaces*, *29*(3), 14–26.
13. Kantor, J., & Maital, S. (1999). Measuring efficiency by product group: Integrating DEA with activity-based accounting in a large Mideast bank. *Interfaces*, *29*(3), 27–36.
14. Leightner, J., & Lovell, C. (1998). The impact of financial liberalisation on the performance of Thai banks. *Journal of Economics and Business*, *50*, 115–131.
15. Mukherjee, A., Nath, P., & Pal, M. N. (2002). Performance benchmarking and strategic homogeneity of Indian banks. *International Journal of Bank Marketing*, *20*(3), 122–139.
16. Pastor, J., & Ruiz, J. (2007). Variables with negative values in DEA. In J. Zhu & W. Cook (Eds.), *Modeling data irregularities and structural complexities in data envelopment analysis* (pp. 63–84). Berlin: Springer.
17. Portela, M., Thanassoulis, E., & Simpson, G. (2004). A directional distance approach to deal with negative data in DEA. *Journal of Operational Research Society*, *55*(10), 111–1121.
18. Saha, A., & Ravishankar, T. S. (2000). Rating of Indian commercial banks: A DEA approach. *European Journal of Operations Research*, *124*, 187–203.
19. Sathye, M. (2001). X-efficiency in Australian banking: An empirical investigation. *Journal of Banking and Finance*, *25*, 613–630.
20. Sathye, M. (2002). Measuring productivity changes in Australian banking: An application of Malmquist indices. *Managerial Finance*, *28*, 48–59.
21. Sherman, D. H., & Gold, F. (1985). Bank branch operating efficiency: Evaluation with data envelopment analysis. *Journal of Banking and Finance*, *9*(3), 297–315.
22. Sturm, J., Williams, B. (2002). Deregulation, entry of foreign banks and bank efficiency in Australia. CESifo Working Paper.
23. Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*(3), 498–509.

The Inventory Management of Fresh Vegetables Using Inventory Balance and Across of Its Supply Chain

Adi Djoko Guritno, Henry Yuliando and Endy Suwondo

Abstract A wider application of the Integrated Planning Support Framework (IPSF) model can provide more practical basis for benchmarking of real case studies. This study is aimed to determine the opportunity loss and cost of excess inventory of fresh vegetables, the optimal inventory level, and the inventory management based on the types of distribution lead-time. An extensive observation over six-month supply chain tracking from several point of sales of fresh vegetables was conducted to employee the proposed IPSF. The result showed an inefficient supply chain due to an imbalance of customer service and inventory levels. Each tier within the supply chain indicated a dissimilar of inventory size that influenced the inventory decision of each tier. In further, it suggests that the losses of fresh vegetables in term of cost and quality could be reduced by using inventory balance and policies in levelling inventory.

1 Background

Supply Chain Management (SCM) deals with a management of business activities in order to obtain raw materials, transformation (processing), a work in process, inventory of finished product, and the distribution system in delivering the products to the consumers [6]. One of the main objectives of SCM is to ensure that the product is available at the right place and the right time to meet consumer demand without creating excess or shortage of stock (inventory).

A. D. Guritno (✉) · H. Yuliando · E. Suwondo
Department of Agroindustrial Technology, Gadjah Mada University, Yogyakarta, Indonesia
e-mail: adidjoko@tip-ugm.org

H. Yuliando
e-mail: henry@tip-ugm.org

E. Suwondo
e-mail: endys@gadjahmada.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_24,

Fujianti et al. [3] describe that among the important factors of supply chain performance are supply chain costs and employed capital. For the fresh vegetable products, a guaranntee that the products purchased by the next tier (middleman) in its supply chain is necessary due to the deterioration of the product that cause a salvage. Perishability of fresh vegetables in term of quality and quantity have confirmed by Kusumaputri et al. [7]. Therefore, it is necessary to manage the flow of product appropriately by determining the best service level. The lack or excess supply cause a negative impact on supply chain performance.

Concerning to a need for inventory management of fresh vegetable product in accordance to the lack of treatment facilities owned by farmers, a trade off between understock versus overstock situation should be prompted. Important to note that the inventory management decisions will affect the performance of the supplier. However, in managing the flow of materials/products at point to point in the supply chain is constrained to several aspects. According to Blos et al. [1], the issue of Supply Chain Risk Management (SCRM) includes two factors e.g. (1) supply chain risks (e.g. operational risks or disruption risks), and (2) mitigation approaches (e.g. supply management, demand management, product management, or information management). Before selecting a proper strategy of supply chain, it is necessary to determine previously which part of the network that is perceived facing an uncertainty of demand, and what policy to anticipate such uncertainty [2].

When farmers or suppliers of fresh vegetables have three channels to deliver their yield either to the supermarkets, traditional markets and restaurants, definitely they are facing an uncertainty regarding to the demand level of each channel. Here, the stock volume affects the barganing position. Therefore, it necessaries to suggest a decision regarding to inventory level inventory balancing of those farmers.

2 Materials and Method

This study was based on a survey over fresh vegetable supply chain in Yogyakarta Province, Indonesia. Four grouped of suppliers were classified e.g. supplier A (those who supplies supermarkets), supplier B as a competitor of supplier A, supplier C as a wholesaler supplier, and supplier D as a restaurant supplier. The product supplied includes several fresh vegetables such as bitroot, broccoli, green beans, radish, chives, spring onions, cauliflower, cabbage, and so on.

The analysis includes both qualitatively and quantitatively. At first, the type of inventory is determined on a base of demand characteristics. Incorporated by the analyis of a decision on how much the inventory should be provided. It is approached by plotting each supplier according to his interest in the supply chain network. Those suppliers could have a specific goals regarding to how are they treat the inventory. Combined with their storage system, the inventory could be aimed for speculation, postponement, consignment or reverse consignment.

Inventory balancing is associated with whether the suppliers have an excess inventory or even a stockout. The formula to find the balance is based on

$SL = Cs/(Cs + Ce)$, where SL is service level, Cs is the cost of having stock-out and Ce is the cost of having excess inventory. Optimal SL in percentage is SL with minimum cost. When SL is found, the Z value (normal distribution value) is used to calculate the amount (Q) of stock level. This value (Q) is in comply with $P\ [D \leq Q] = SL$. Since it can be always assumed normally distributed then the condition is valid for $D \sim N(\mu D,\ \sigma D)$. To calcultate the inventory balance is as follows:

$$\text{Inventory Balance} = \mu D + zSL \times \sigma D \tag{1}$$

where:

μD demand during lead time
zSL z value times SL
σD demand standard deviation.

The common factors underlying the selection of decision-making in inventory management includes the purpose of usage, the nature of supply chain, and bargaining power. These three factors should be considered carefully when choosing the right approach to inventory management. Particularly for the lead time factor, it can be divided into several aspects in comply with customer order to fulfillment lead time (CLT), supplier order to fulfillment lead time (SLT), cycle time (CT), and delivery to customer lead time (DTC). The combination of these aspects against lead time lead to an option that aiming to inventory management. Improving transformation process and shortened lead time are key activities in bioproduction system. Bioproduction system is identical to the development of plant factory. It involves technologies such as process control for the plant growth environment, mechanization for material handling, system control for production and computer applications [5]. Wallin et al. [10] has proposed a tabulation that can lead to the right decision in inventory management as presented in Fig. 1.

3 Result and Discussion

Research was conducted in the central of vegetable production areas in Central Java, and the market areas spread out in southern part of Central Java and Yogyakarta province. The source of fresh vegetables from group of farmers have two distinctive order and harvetsing types, namely OBP (Order Before Planting) and OBH (Order Before Harvesting) with different impact of transportation aproach used. Fresh vegetables with OBP approach tend to use LTL (Less Than Truckload) and OBH use TL (Truck Load) transportation [4]. Description of supply chain stages captured by IPSF shows some stages flow of fresh vegetables from the sources area to the point of sales that consist of 4 arrangements: order type and transportation, distribution and market, cost of inventory and order lead time, and inventory practices arrangement in each stage across supply chain (Fig. 2). Based on the analysis, a summary of the results are presented as follows. Based on the tables, it can noted that supplier A has

Factors in inventory management			
Customer need	Supply Chain Character	Bargaining position	Inventory Approach
CLT < SLT + CT + DTC	Unreliable supply chain	Choose several suppliers	Speculation Inventory
Difficult to predict demand and preference	Unpredictable the performance of order and delivery	Supplier provide a unique product	
Stable customer preference			
CLT > SLT + CT + DTC	Unreliable supply chain	Choose many suppliers	Postponement Inventory
Easy to predict demand and preference	Predictable the performance of order and delivery	Supplier provide non-unique product	
Quick change in customer preference			
CLT < SLT + CT + DTC	Unreliable supply chain	Choose many suppliers	Consignment Inventory
Difficult to predict demand and preference	Unpredictable the performance of order and delivery	Supplier provide non-unique product	
Quick change in customer preference			
CLT > SLT + CT + DTC	Reliable supply chain	Choose many suppliers	Reserve Inventory Consigment
Easy to predict demand and preference	Predictable the performance of order and delivery	Supplier provide unique product	
Stable customer preference			

Fig. 1 Factors that influencing the decision in inventory management

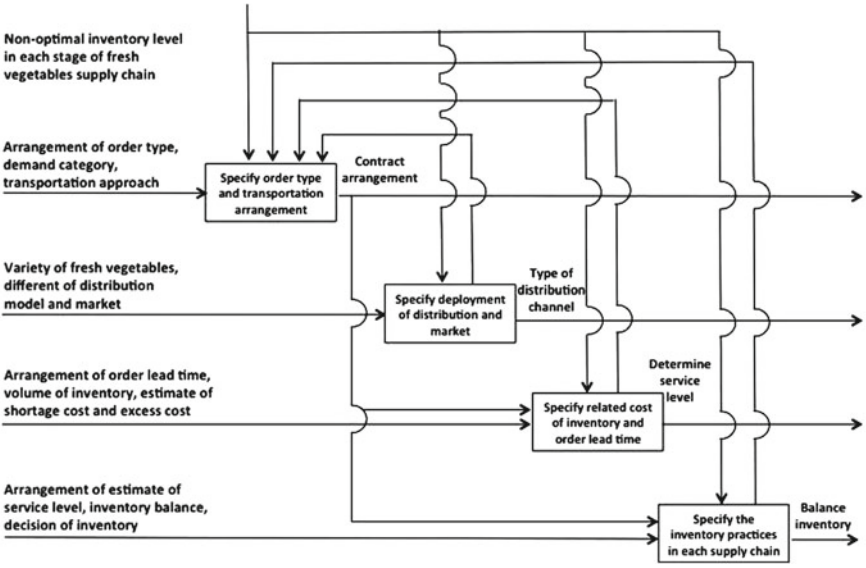


Fig. 2 IPSF model for fresh vegetable supply chain

an average demand level and inventory balancing that significant. This indicates that their demand prompts a high fluctuation. When the stockout occured, the inventory balancing provide an advantages due to the product availability, and vice versa. It can be analyzed that the excess demand is projected to be above of the level of safety stock at a particular time. A shortage also occured at several products means that the opportunity losses of profit will also occur. Theoretically, insurance is an option of risk mitigation by definition. But other way out can be used, such as to include information sharing schemes [8, 9] (Tables 1, 2, 3, 4).

Table 1 Inventory balancing for supplier A

Stock keeping unit	*Cs* (Rupiah)	*Ce* (Rupiah)	SL	*z* value	μ demand (kg)	σ demand (kg)	Inv balance (kg)
Cabbage	1,000	228	0.8143	0.89	96.98	30.53	124.15
Brocoli	2,000	325	0.8602	1.08	38.46	29.02	69.80
Chicory	1,000	293	0.7734	0.62	43.28	43.82	70.45
Tomato	1,000	163	0.8598	1.08	32.75	19.91	54.26
Squash	1,000	65	0.9390	1.54	32.87	20.44	64.35
Pakcoy	1,000	130	0.8850	1.20	15.42	8.52	25.64
Lettuce	2,000	325	0.8602	1.08	16.84	24.54	43.34
Spring onion	1,000	163	0.8598	1.08	7.06	4.51	11.93
Celery	2,000	195	0.9112	1.35	7.48	2.33	10.63
Green lettuce	2,000	260	0.8850	1.20	7.40	5.79	14.35
Kucay	2,000	163	0.9246	1.43	03.18	1.23	4.94
Parsley	2,000	325	0.8602	1.08	2.40	0.56	03.01

Note US$ 1 = Rp 11,500.00

Table 2 Inventory balancing for supplier B

Stock keeping unit	*Cs* (Rupiah)	*Ce* (Rupiah)	SL	*z* value	μ demand (kg)	σ demand (kg)	Inv balance (kg)
Spinach	2,000	195	0.9122	1.34	26.60	0.29	26.99
Kale	2,500	244	0.9111	1.34	17.60	1.44	19.53
Caysim	2,000	260	0.8850	1.20	17.16	3.40	21.24
Spring onion	2,000	260	0.8850	1.20	9.05	2.85	12.47
Green lettuce	3,000	423	0.8764	1.15	18.60	1.53	20.36
Bean	3,000	309	0.9066	1.32	79.50	34.77	125.40
Tomato	3,500	423	0.8922	1.23	15.60	2.67	18.89
Brocoli	4,000	488	0.8913	1.22	53.00	10.58	65.91
Red spinach	3,000	390	0.8850	1.20	16.90	1.23	18.38
Kailan	2,500	325	0.8850	1.20	31.16	0.50	31.76

Note US$ 1 = Rp 11,500.00

Table 3 Inventory balancing for supplier C

Stock keeping unit	*Cs* (Rupiah)	*Ce* (Rupiah)	SL	*z* value	μ demand (kg)	σ demand (kg)	Inv balance (kg)
Brocoli	1,300	342	0.7917	0.81	55.50	0.35	55.79
Pakcoy	800	147	0.8448	1.11	35.40	0.50	35.96
Cabbage	800	244	0.7663	0.72	33.80	0.29	34.01
Bean	1,000	33	0.9681	1.86	22.80	1.04	24.74
Bitroot	2,000	390	0.8368	0.98	21.80	1.04	22.82
Green lettuce	1,200	293	0.8038	0.85	16.10	0.29	16.35
Spring onion	800	179	0.8172	0.90	10.10	0.20	10.36

Note US$ 1 = Rp 11,500.00

Table 4 Inventory balancing for supplier D

Stock keeping unit	*Cs* (Rupiah)	*Ce* (Rupiah)	SL	z value	μ demand (kg)	σ demand (kg)	Inv balance (kg)
Lettuce	1,100	358	0.7545	0.68	18.00	1.00	18.68
Bean	900	228	0.7979	0.83	16.00	0.50	16.42
Spring onion	1,200	195	0.8602	1.80	4.60	0.36	5.25
Celery	1,500	228	0.8681	1.11	4.20	0.29	4.52
Green lettuce	1,500	293	0.8366	0.98	1.60	0.36	1.95

Note US$ 1 = Rp 11,500.00

In contrast, the inventory balancing of supplier B, C and D behaved oppositely. The resulting inventory balancing was not significant. This indicates that demand was fairly constant. This indicates also that suppliers did not need to have an excess inventory to anticipate the product shortages. In this case, each supplier has different inventory balancing. Refer to the expected inventory balancing, the supplier can optimize the return of their sale. Thus, this is an optimal point where the supplier can earn optimum profit and optimum service to consumers. In further, optimal service to consumers determine the service level.

4 Conclusion

The given lead time for all fresh vegetable products of each supplier studied here are relatively short, but few commodities. This proved that in general, the nature of the supply chain of each supplier can be reliable, both in quantity and timely delivery. Suppliers (farmers/middleman/distributor) who supply both types of non-unique and unique products obtain a strong bargaining position. Inventory decisions that fit to this case study are aiming for speculation and postponement. Inventory balance of suppliers B, C and D are significant compared to the average demand available for supplier A who were facing the fluctuation demand. The results showed no significant differences regarding to the demand characteristics or can assumed relatively constant.

References

1. Blos, M. F., Quaddus, M., & Wee, H. M. (2009). Supply chain risk management: A case study on the automotive and electronic industries in Brazil. *Supply Chain Management: An International Journal*, *14*(4), 247–252.
2. Cucchiella, F., & Gastaldi, M. (2006). Risk management in supply chain: A real option approach. *Journal of Manufacturing Technology Management*, *17*(6), 700–720.
3. Fujianti, R., Guritno, A. D., Suwondo, E. (2012). Valuation of supply chain performance in fresh vegetables using the analytical hierarchy process (AHP) and supply chain operations references

(SCOR). In Proceedings of a conference on the development of national competitiveness (434 p.), Nov 16–18, 2011, Yogyakarta.
4. Guritno, A. D. (2013). Development of supply chain risk management of fresh vegetables. In Proceeding of food innovation Asia conference 2013: Empowering SMEs through science and technology, Bangkok, Thailand.
5. Guritno, A. D., Suwondo, E., Yuliando, H., Ushada, M., & Murase, H. (2013). Development of drum-buffer-rope algorithm to control capacity constrained machine in a bioproduction system. In Proceeding of the 2013 IFAC bio-robotics conference, International Federation of Automatic Control, Osaka, Japan.
6. Heizer, J., & Render, B. (2010). *Operation management* (10th ed.). New York: Prentice Hall.
7. Kusumaputri, D. A., Guritno, A. D., Suwondo, E. (2012). Evaluation of inventory decision and inventory balance of fresh vegetables in several stages of supply chains. In Proceedings of a conference on the development of national competitiveness (434 p.), Nov 16–18, 2011, Yogyakarta.
8. Olson, D. L., & Wu, D. (2011). Risk management models for supply chain: A scenario analysis of outsourcing to China. *Supply Chain Management: An International Journal*, *16*(6), 401–408.
9. Punniyamoorthy, M., Thamasaiselvan, N., & Manikandan, L. (2011). Assessment of supply chain risk: Scale development and validation. *Benchmarking: An International Journal*, *20*(1), 79–105.
10. Wallin, C., Rungtusanatham, M. J., & Rabinovich, E. (2006). What is the right inventory management approach for a purchased item? *International Journal of Operations & Production Management*, *26*(1), 50–68.

Moving Bins from Conveyor Belts onto Pallets Using FIFO Queues

Frank Gurski, Jochen Rethmann and Egon Wanke

Abstract We study the combinatorial FIFO stack-up problem. In delivery industry, bins have to be stacked-up from conveyor belts onto pallets. Given k sequences $q_1, \ldots, q_k$ of labeled bins and a positive integer p, the goal is to stack-up the bins by iteratively removing the first bin of one of the k sequences and put it onto a pallet located at one of p stack-up places. Each of these pallets has to contain bins of only one label, bins of different labels have to be placed on different pallets. After all bins of one label have been removed from the given sequences, the corresponding place becomes available for a pallet of bins of another label. The FIFO stack-up problem is NP-complete in general. In this paper we show that the problem can be solved in polynomial time, if the number k of given sequences is fixed.

1 Introduction

We consider the combinatorial problem of stacking up bins from a set of conveyor belts onto pallets. This problem originally appears in *stack-up systems* that play an important role in delivery industry and warehouses. A detailed description of the practical background of this work is given in [2].

The bins that have to be stacked up onto pallets arrive at the stack-up system on the main conveyor of an order-picking system. At the end of the main conveyor

F. Gurski (✉) · E. Wanke
Institute of Computer Science, Heinrich Heine University, 40225 Düsseldorf, Germany
e-mail: frank.gurski@hhu.de

E. Wanke
e-mail: e.wanke@hhu.de

J. Rethmann
Faculty of Electrical Engineering and Computer Science, Niederrhein
University of Applied Sciences, 47805 Krefeld, Germany
e-mail: jochen.rethmann@hs-niederrhein.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_25,

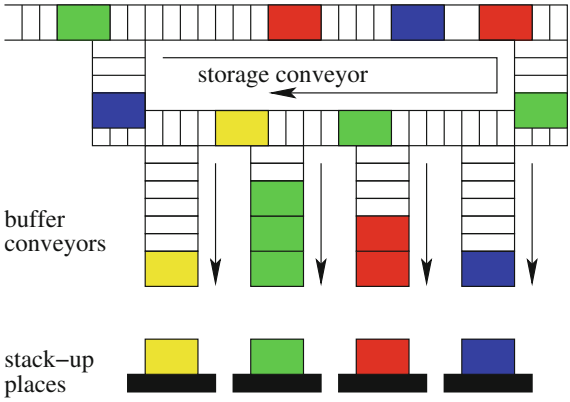


Fig. 1 A real stack-up system

they enter a *cyclic storage conveyor*, see Fig. 1. From the storage conveyor the bins are pushed out to *buffer conveyors*, where they are queued. The bins are picked-up by stacker cranes from the end of a buffer conveyor and moved onto pallets, which are located at some *stack-up places*. There is one buffer conveyor for each stack-up place. Automatic driven vehicles take full pallets from stack-up places, put them onto trucks and bring new empty pallets to the stack-up places.

In real life, the cyclic storage conveyor is necessary to enable a smooth stack-up process irrespective of the real speed the cranes and conveyors are moving. Such details are unnecessary to compute an order in which the bins can be palletized. So, in our model we neglect the cyclic storage conveyor, and the number of stack-up places is not correlated to the number of sequences. Logistic experiences over 30 years lead to such high flexible conveyor systems in delivery industry. So, we do not intend to modify the architecture of existing systems, but try to develop efficient algorithms to control them. Figure 2 shows a sketch of a simplified stack-up system with 2 buffer conveyors and 3 stack-up places.

The FIFO stack-up problem has important practical applications. Many facts are known on stack-up systems that use a random access storage instead of buffer queues. The stack-up problem with random access storage is NP-complete, but can be solved efficiently if the storage capacity s or the number of stack-up places p is fixed. It remains NP-complete as shown in [5], even if the sequence contains at most 9 bins per pallet. A polynomial time *off-line* approximation algorithm is introduced in [5] that yields a processing that is optimal up to a factor bounded by $\log(p)$. In [6] the performances of some simple *on-line* stack-up algorithms are compared with optimal off-line solutions by a competitive analysis [1, 4].

The FIFO stack-up problem is NP-complete even if the number of bins per pallet is bounded [3]. In this paper we show by dynamic programming that the FIFO stack-up problem can be solved in polynomial time for a fixed number k of sequences.

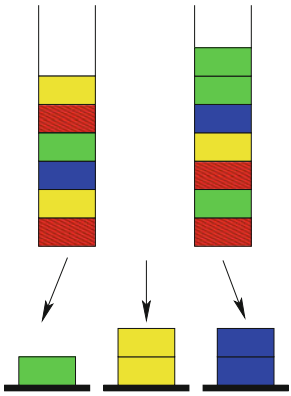


Fig. 2 A FIFO stack-up system

2 Preliminaries

Let k and p be two positive integers. We consider *sequences* $q_1 = (b_1, \ldots, b_{n_1}), \ldots, q_k = (b_{n_{k-1}+1}, \ldots, b_{n_k})$ of pairwise distinct *bins*. These sequences represent the buffer queues in real stack-up systems. Each bin b is labeled with a *pallet symbol* $plt(b)$. We say bin b is destined for pallet $plt(b)$. In examples we often choose characters as pallet symbols. The set of all pallets of the bins in some sequence q_i is denoted by $plts(q_i) = \{plt(b) \mid b \in q_i\}$. For a list of sequences $Q = (q_1, \ldots, q_k)$ we denote $plts(Q) = plts(q_1) \cup \cdots \cup plts(q_k)$. For some sequence $q = (b_1, \ldots, b_n)$ we say bin b_i is *on the left of* bin b_j in sequence q if $i < j$. A sequence $q' = (b_j, b_{j+1}, \ldots, b_n)$, $j \geq 1$, is called a *subsequence* of sequence $q = (b_1, \ldots, b_n)$, and we write $q - q' = (b_1, \ldots, b_{j-1})$.

Let $Q = (q_1, \ldots, q_k)$ and $Q' = (q'_1, \ldots, q'_k)$ be two lists of sequences of bins, such that each sequence q'_j, $1 \leq j \leq k$, is a subsequence of sequence q_j. Each such pair (Q, Q') is called a *configuration*. A pallet t is called *open* in configuration (Q, Q'), if a bin of pallet t is contained in some $q'_i \in Q'$ and if another bin of pallet t is contained in some $q_j - q'_j$ for $q_j \in Q$, $q'_j \in Q'$. The *set of open pallets* in configuration (Q, Q') is denoted by $open(Q, Q')$. A pallet $t \in plts(Q)$ is called *closed* in configuration (Q, Q'), if $t \notin plts(Q')$, i.e. no sequence of Q' contains a bin for pallet t.

3 The FIFO Stack-up Problem

Consider a configuration (Q, Q'). The removal of the first bin from one subsequence $q' \in Q'$ is called *transformation step*. A sequence of transformation steps that transforms the list Q into empty subsequences is called a *processing* of Q.

Given a list $Q = (q_1, \ldots, q_k)$ of sequences and a positive integer p, the *FIFO stack-up problem* is to decide whether there is a processing of Q, such that in each configuration (Q, Q') during the processing at most p pallets are open.

It is often convenient to use pallet identifications instead of bin identifications to represent a sequence q. For n not necessarily distinct pallets $t_1, \ldots, t_n$ let $[t_1, \ldots, t_n]$ denote some sequence of n pairwise distinct bins $(b_1, \ldots, b_n)$, such that $plt(b_i) = t_i$ for $i = 1, \ldots, n$. We use this notion for lists of sequences as well. For the sequences $q_1 = [t_1, \ldots, t_{n_1}], \ldots, q_k = [t_{n_{k-1}+1}, \ldots, t_{n_k}]$ of pallets we define $q_1 = (b_1, \ldots, b_{n_1}), \ldots, q_k = (b_{n_{k-1}+1}, \ldots, b_{n_k})$ to be sequences of bins such that $plt(b_i) = t_i$ for $i = 1, \ldots, n_k$, and all bins are pairwise distinct.

Consider a processing of a list Q of sequences. Let $B = (b_{\pi(1)}, \ldots, b_{\pi(n)})$ be the order in which the bins are removed during the processing of Q, and let $T = (t_1, \ldots, t_m)$ be the order in which the pallets are opened during the processing of Q. We call B a *bin solution* of Q, and T is called a *pallet solution* of Q.

During a processing of a list Q of sequences there are often configurations (Q, Q') for which it is easy to find a bin b that can be removed from Q' such that a further processing with p stack-up places is still possible. This is the case, if bin b is destined for an already open pallet. Consider a processing of some list Q of sequences with p stack-up places. Let $(b_{\pi(1)}, \ldots, b_{\pi(i-1)}, b_{\pi(i)}, \ldots, b_{\pi(l-1)}, b_{\pi(l)}, b_{\pi(l+1)}, \ldots, b_{\pi(n)})$ be the order in which the bins are removed from the sequences during the processing, and let (Q, Q_i), $1 \leq i \leq n$ denote the configuration such that bin $b_{\pi(i)}$ is removed in the next transformation step. Suppose bin $b_{\pi(i)}$ will be removed in some transformation step although bin $b_{\pi(l)}$, $l > i$, for some already open pallet $plt(b_{\pi(l)}) \in open(Q, Q_i)$ could be removed next. We define a modified processing $(b_{\pi(1)}, \ldots, b_{\pi(i-1)}, b_{\pi(l)}, b_{\pi(i)}, \ldots, b_{\pi(l-1)}, b_{\pi(l+1)}, \ldots, b_{\pi(n)})$ by first removing bin $b_{\pi(l)}$, and afterwards the bins $b_{\pi(i)}, \ldots, b_{\pi(l-1)}$ in the given order. Obviously, in each configuration during the modified processing there are at most p pallets open. A configuration (Q, Q') is called a *decision configuration*, if the first bin of each sequence $q' \in Q'$ is destined for a non-open pallet. FIFO stack-up algorithms will only be asked for a decision in such decision configurations, in all other configurations the algorithm automatically removes a bin for some already open pallet.

If we have a pallet solution computed by any FIFO stack-up algorithm, we can convert the pallet solution into a sequence of transformation steps, i.e. a processing of Q by some simple algorithm not shown here because of space restrictions [3].

4 Main Result

Our aim in controlling FIFO stack-up systems is to compute a processing of the given sequences of bins with a minimum number of stack-up places. Such an optimal processing can always be found by computing the *processing graph* and doing some calculation on it. Before we define the processing graph let us consider some general graph problem, that is strongly related to the FIFO stack-up problem.

Let $G = (V, E, f)$ be a directed acyclic vertex-labeled graph. Function $f: V \to \mathbb{Z}$ assigns to every vertex $v \in V$ a value $f(v)$. Let $s \in V$ and $t \in V$ be two vertices. For some vertex $v \in V$ and some path $P = (v_1, \ldots, v_\ell)$ with $v_1 = s$, $v_\ell = v$ and $(v_i, v_{i+1}) \in E$ we define $val_P(v) := \max_{u \in P}(f(u))$. Let $\mathscr{P}_s(v)$ denote the set of all paths from vertex s to vertex v. We define $val(v) := \min_{P \in \mathscr{P}_s(v)}(val_P(v))$. A solution of this problem can be found by dynamic programming and solves also the FIFO stack-up problem. For some vertex $v \in V$ let $N^-(v) := \{u \in V \mid (u, v) \in E\}$ be the set of predecessors of v in graph G. Then it holds: $val(v) = \max\{f(v), \min_{u \in N^-(v)}(val(u))\}$. If we compute the value of $val(v)$ recursively, subproblems often would be calculated several times. So, we use dynamic programming to calculate each subproblem only once, and put them together by already calculated sub-solutions. This is possible, since the graph is directed and acyclic.

Let $topol: V \to \mathbb{N}$ be a topological ordering of the vertices of the graph G, i.e. an ordering of the vertices such that $topol(u) < topol(v)$ holds for every $(u, v) \in E$.

The value $val(t)$ can be computed in polynomial time. We need some additional terms, before we show how the above procedure can solve the FIFO stack-up problem. For a sequence $q = (b_1, \ldots, b_n)$ let $left(q, i) := (b_1, \ldots, b_i)$ denote the sequence of the first i bins of sequence q, and let $right(q, i) := (b_{i+1}, \ldots, b_n)$ denote the remaining bins of sequence q after removing the first i bins. It can be seen that a configuration is well-defined by the number of bins that are removed from each sequence. The position of the first bin in some sequence q_i destined for some pallet t is denoted by $first(q_i, t)$, similarly the position of the last bin for pallet t in sequence q_i is denoted by $last(q_i, t)$.

Example 1 Consider list $Q = (q_1, q_2)$ of the sequences $q_1 = [a, b, c, a, b, c]$ and $q_2 = [d, e, f, d, e, f, a, b, c]$. Then we get $left(q_1, 2) = [a, b]$, $right(q_1, 2) = [c, a, b, c]$, $left(q_2, 3) = [d, e, f]$, and $right(q_2, 3) = [d, e, f, a, b, c]$.

If we denote $q_1' := right(q_1, 2)$, and $q_2' := right(q_2, 3)$, then, in Example 1, there are 5 pallets open in configuration (Q, Q') with $Q' = (q_1', q_2') : a, b, d, e$, and f. We generalize this for a list $Q = (q_1, \ldots, q_k)$ of sequences and we define the cut $cut_Q(i_1, \ldots, i_k) := \{t \in plts(Q) \mid \exists j, j', b \in left(q_j, i_j), b' \in right(q_{j'}, i_{j'}): plt(b) = plt(b') = t\}$ at some configuration $(i_1, \ldots, i_k)$ to be the set of pallets t such that one bin for pallet t has already been removed and another bin for pallet t is still contained in some sequence. Let $\#cut_Q(i_1, \ldots, i_k)$ be the number of elements in $cut_Q(i_1, \ldots, i_k)$.

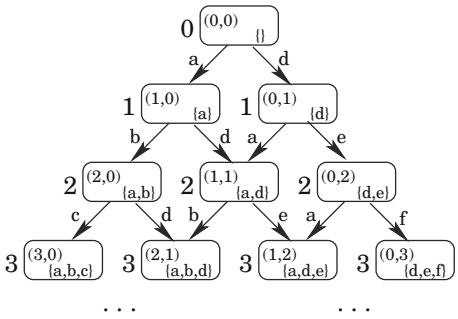
The intention of a processing graph $G = (V, E, h)$ is the following. Suppose each vertex $v \in V$ of graph G represents a configuration $(i_1, \ldots, i_k)$ during the processing of some set of sequences Q. Suppose further, an edge $(u, v) \in E$ represents a transformation step during this processing, such that a bin b is removed from some sequence in configuration u resulting in configuration v. Suppose also that each vertex v is assigned the number of open pallets in configuration v, i.e. number $\#cut_Q(i_1, \ldots, i_k)$. If vertex s represents the initial configuration $(0, 0, \ldots, 0)$, while vertex t represents the final configuration $(|q_1|, |q_2|, \ldots, |q_k|)$, then we are searching

Fig. 3 Dynamic programming

Algorithm OPT

$val[s] := f(s)$
for every $v \neq s$ in order of *topol* do
  $val[v] := \infty$
  for every $u \in N^-(v)$ do
    if($val[u] < val[v]$)
      $val[v] := val[u]$
      $pred[v] := u$
  if ($val[v] < f(v)$)
    $val[v] := f(v)$

Fig. 4 The processing graph of Example 1



a path P from s to t such that the maximal number on path P is minimal. Thus, an optimal processing of Q can be found by Algorithm OPT given in Fig. 3.

The processing graph has a vertex for each possible configuration. Each vertex v of the processing graph is labeled by the vector $h(v) = (v_1, \ldots, v_k)$, where v_i denotes the position of the bin that has been removed last from sequence q_i. There is a directed edge from vertex u labeled by $(u_1, \ldots, u_k)$ to vertex v labeled by $(v_1, \ldots, v_k)$ if and only if $u_i = v_i - 1$ for exactly one element of the vector and for all other elements of the vector $u_j = v_j$. The edge is labeled with the pallet symbol of that bin, that will be removed in the corresponding transformation step. For the sequences of Example 1 we get the processing graph of Fig. 4. The processing graph is directed and acyclic, and we use this graph to compute the values of $\#cut_Q(i_1, \ldots, i_k)$ iteratively in the following way.

First, since none of the bins has been removed from any sequence, we have $\#cut_Q(0, \ldots, 0) = 0$. Since the processing graph is directed and acyclic, there exists a topological ordering $topol$ of the vertices. The vertices are processed according to the order $topol$. In each transformation step we remove exactly one bin for some pallet t from some sequence q_j, thus

$$\#cut_Q(i_1, \ldots, i_{j-1}, i_j, i_{j+1}, \ldots, i_k) = \#cut_Q(i_1, \ldots, i_{j-1}, i_j - 1, i_{j+1}, \ldots, i_k) + c_j$$

Table 1 Additional hash-tables to perform all operations efficiently

Pallet t	a	b	c	d	e	f
$first(q_1, t)$	1	2	3	.	.	.
$first(q_2, t)$	7	8	9	1	2	3
$last(q_1, t)$	4	5	6	.	.	.
$last(q_2, t)$	7	8	9	4	5	6

where

$$c_j = \begin{cases} 1, \text{ if } first(q_j, t) = i_j \text{ and } (t \notin plts(q_\ell) \text{ or } first(q_\ell, t) > i_\ell) \ \forall\, \ell \neq j \\ -1, \text{ if } last(q_j, t) = i_j \text{ and } (t \notin plts(q_\ell) \text{ or } last(q_\ell, t) \leq i_\ell) \ \forall\, \ell \neq j \\ 0, \text{ otherwise.} \end{cases}$$

That means, $c_j = 1$ if pallet t has been opened in the last transformation step, and $c_j = -1$ if pallet t has been closed in the last transformation step. Otherwise, c_j is zero. Thus, the calculation of value $\#cut_Q(i_1, \ldots, i_k)$ for the vertex labeled $(i_1, \ldots, i_k)$ depends only on already calculated values. Figure 4 shows such a processing for the sequences of Example 1. To efficiently perform this processing, we have to store for each pallet the first and last bin in each sequence. We use hash-tables to efficiently store such values without the need of initializing the values of absent pallets. Table 1 shows such hash-tables for the sequences of Example 1.

The calculation of those tables can be done in time $\mathcal{O}(|q_1| + \cdots + |q_k|) = \mathcal{O}(k \cdot \max_{1 \leq i \leq k} |q_i|) = \mathcal{O}(\max_{1 \leq i \leq k} |q_i|)$, since k is fixed. Afterwards, the computation of each value c_j can be done in time $\mathcal{O}(k)$. After the computation of each value $\#cut_Q(i_1, \ldots, i_k)$, we can use Algorithm OPT to compute the minimal number of stack-up places necessary to process the given FIFO stack-up problem. If the size of the processing graph is polynomial bounded in the size of the input, the FIFO stack-up problem can be solved in polynomial time.

References

1. Borodin, A. (1998). *On-line computation and competitive analysis*. Cambridge: Cambridge University Press.
2. de Koster, R. (1994). Performance approximation of pick-to-belt orderpicking systems. *European Journal of Operational Research*, *92*, 558–573.
3. Gurski, F., Rethmann, J., & Wanke, E. (2013). Complexity of the fifo stack-up problem. *ACM Computing Research Repository (CoRR)*, abs/1307.1915.
4. Manasse, M. S., Mc Geoch, L. A., & Sleator, D. D. (1988). Competitive algorithms for on-line problems. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, ACM, pp. 322–333.
5. Rethmann, J., & Wanke, E. (2000). On approximation algorithms for the stack-up problem. *Mathematical Methods of Operations Research*, *51*, 203–233.
6. Rethmann, J., & Wanke, E. (2001). Stack-up algorithms for palletizing at delivery industry. *European Journal of Operational Research*, *128*(1), 74–97.

A Generalization of Odd Set Inequalities for the Set Packing Problem

Olga Heismann and Ralf Borndörfer

Abstract The set packing problem, sometimes also called the stable set problem, is a well-known NP-hard problem in combinatorial optimization with a wide range of applications and an interesting polyhedral structure, that has been the subject of intensive study. We contribute to this field by showing how, employing cliques, odd set inequalities for the matching problem can be generalized to valid inequalities for the set packing polytope with a clear combinatorial meaning.

1 Introduction and Terminology

The set packing problem, sometimes also called the stable set problem, is a well-known NP-hard [4] problem in combinatorial optimization with a wide range of applications. Its weighted version can be formulated as follows. Given a finite set V and some set $E \subseteq 2^V$ with weights assigned to each set in E, find a subset of pairwise disjoint sets from E, called a packing, with a maximum sum of weights. Although many classes of facets for the set packing problem polytope are known (see, e.g., [1]), there is still no complete polyhedral description known and further facet classes have to be researched.

A polynomially solvable special case of the set packing problem, where all sets in E have size two, is the matching problem. For this problem, the polytope can be completely described by adding so-called odd set inequalities to the canonical description [3]. In this paper, we show how, employing cliques, the odd set inequalities can be generalized to valid inequalities for the set packing problem polytope with a clear combinatorial meaning. For the hypergraph assignment problem [2], a partitioning

O. Heismann (✉) · R. Borndörfer
Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany
e-mail: heismann@zib.de

R. Borndörfer
e-mail: borndoerfer@zib.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_26,

problem on bipartite hypergraphs, inequalities from this class can be facet-defining. We also relate the presented inequality class to a different generalization of odd set inequalities for the stable set problem called general clique family inequalities [5].

After summarizing the basic terminology needed in what follows, we present a combinatorial derivation of the inequality class. In the end, we show a comparison with the general clique family inequalities.

Definition 1 A *hypergraph* $G = (V, E)$ is a pair of a *vertex* set V and a set $E \subseteq 2^V \setminus \emptyset$ of subsets of V called *hyperedges*. A *packing* $H \subseteq E$ in G is a subset of pairwise disjoint hyperedges, i. e., $e_1 \cap e_2 = \emptyset$ for all $e_1, e_2 \in H$ with $e_1 \neq e_2$.

If all hyperedges have size k, i.e., $|e| = k$ for all $e \in E$, G is called *k-uniform*. A two element hyperedge is also called an *edge*. If all hyperedges are edges, i.e., the hypergraph G is 2-uniform, G is also called a *graph*.

The set packing problem can then be stated as follows:

Problem 1 (*Set Packing Problem*)
Input: A pair (G, c_E) consisting of a hypergraph $G = (V, E)$ and a cost function $c_E \colon E \to \mathbb{R}$.
Output: A maximum cost packing in G w.r.t. c_E, i.e., a packing H^* in G such that

$$\sum_{e \in H^*} c_E(e) = \max \left\{ \sum_{e \in H} c_E(e) \colon H \text{ is a packing in } G \right\}.$$

The set packing problem can also be formulated as an integer linear program. The canonical formulations is the following.

$$\underset{x \in \mathbb{R}^E}{\text{maximize}} \quad \sum_{e \in E} c_E(e) x_e \tag{SSP}$$

$$\text{subject to} \quad \sum_{e \in \delta(v)} x_e \leq 1 \quad \forall v \in V \tag{i}$$

$$x \geq 0 \tag{ii}$$

$$x \in \mathbb{Z}^E. \tag{iii}$$

Let $P(\text{SSP}) := \text{conv}\{x \in \mathbb{R}^E \colon \text{(SSP) (i)–(iii)}\}$ and $P_{\text{LP}}(\text{SSP}) := \{x \in \mathbb{R}^E \colon \text{(SSP)(i)–(ii)}\}$ be the polytopes associated with the integer linear program (SSP) and its LP relaxation, respectively.

At the end of our generalization procedure, we will substitute vertices by hyperedge cliques. They are defined as follows.

Definition 2 A *hyperedge clique* in a hypergraph $G = (V, E)$ is a set $Q \subseteq E$ of hyperedges such that every two hyperedges $e_1, e_2 \in Q$ have at least one vertex in common, i. e., $e_1 \cap e_2 \neq \emptyset$.

Associated with the hyperedge clique Q is the clique inequality $\sum_{e \in Q} x_e \leq 1$.

2 Generalizing Odd Set Inequalities

Consider the set packing problem for the hypergraph $G = (V, E)$.

In the special case that G is a graph, the set packing problem becomes an edge packing problem which can be completely described by the following system of inequalities [3]:

$$\sum_{e \in \delta(v)} x_e \leq 1 \qquad \forall v \in V \qquad \text{(MP i)}$$

$$\sum_{e \in E : e \subset V'} x_e \leq \frac{|V'| - 1}{2} \quad \forall V' \subseteq V,\ |V'| \geq 3,\ |V'| \text{ odd} \qquad \text{(MP ii)}$$

$$x \geq 0 \qquad \forall e \in E \qquad \text{(MP iii)}$$

The inequalities (MP ii) are called odd set inequalities. Their combinatorial meaning is that for every odd set $V' \subseteq V$ of $|V'| = 2k + 1$ vertices there can be at most $\left\lfloor \frac{|V'|}{2} \right\rfloor = \frac{|V'|-1}{2} = k$ edges connecting pairs of them in a matching. This holds since every edge is incident to two vertices in k, every vertex can be incident to at most one edge in a matching, and $k+1$ edges would need therefore already $2k+2 > |V'|$ distinct vertices.

A formal proof of validity for odd set inequalities can be interpreted as a Chvátal-Gomory procedure with coefficient $\frac{1}{2}$ for all inequalities of type (SSP) (i) for $v \in V'$ and 0 for all others.

We will generalize these inequalities for the set packing problem, i.e., from graphs to hypergraphs, in three steps. The first one will adapt the odd set inequalities to p-uniform hypergraphs, i.e., to hypergraphs which have hyperedges all of size p, where p can be greater than two. Then, we will tackle hypergraphs with hyperedges of arbitrary size by viewing them as combinations of hyperedges of size p in the second step. The third step will generalize sets of hyperedges incident to one vertex to hyperedge cliques.

Odd set inequalities can be also written as

$$\sum_{e \in E} \left\lfloor \frac{|\{v \in V' : e \in \delta(v)\}|}{2} \right\rfloor x_e \leq \left\lfloor \frac{|V'| - 1}{2} \right\rfloor,$$

which is a more useful representation for our generalization procedure.

Step 1. Let G be p-uniform. Applying the idea of odd set inequalities to this situation yields that for every set $V' \subseteq V$ of $|V'| = pk + r$, $0 \leq r \leq p - 1$ vertices there can be at most $\left\lfloor \frac{|V'|}{p} \right\rfloor = \frac{|V'|-r}{p} = k$ hyperedges, each connecting p of them, in a packing. For an example see Fig. 1.

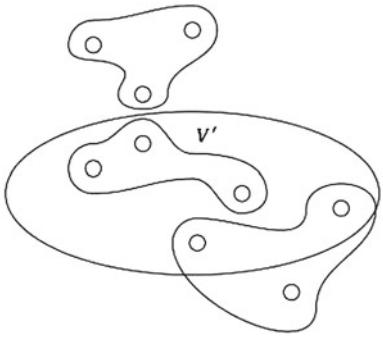


Fig. 1 A packing in a 3-uniform hypergraph $G = (V, E)$ with nine vertices and a vertex subset V' surrounded by an ellipse. There can be at most $\lfloor \frac{5}{3} \rfloor = 1$ hyperedge which is a subset of the five vertices in V'. All other hyperedges in the packing have to have at least one vertex from $V \setminus V'$

This leads to the inequality

$$\sum_{e \in E} \left\lfloor \frac{|\{v \in V' : e \in \delta(v)\}|}{p} \right\rfloor x_e \leq \left\lfloor \frac{|V'|}{p} \right\rfloor \quad \forall V' \subseteq V.$$

The coefficients $\left\lfloor \frac{|\{v \in V' : e \in \delta(v)\}|}{p} \right\rfloor$ all have value 0 or 1.

The inequality can be also derived using a Chvátal-Gomory procedure with coefficient $\frac{1}{p}$ for all inequalities of type (SSP) (i) for $v \in V'$ and 0 for all others.

Step 2. Let G be an arbitrary hypergraph. Choose some $p \in \mathbb{N}$, $p \geq 2$. Contrary to the previous case, where all hyperedges had size p, there now might be hyperedges in the packing that contain more than p vertices from V'. The inequality from Step 1, however, is still true. A hyperedge that contains $kp + r$ vertices from V' can be viewed as k hyperedges of size p that are contained in V'. For an example see Fig. 2.

This idea leads to the inequality class

$$\sum_{e \in E} \left\lfloor \frac{|\{v \in V' : e \in \delta(v)\}|}{p} \right\rfloor x_e \leq \left\lfloor \frac{|V'|}{p} \right\rfloor \quad \forall V' \subseteq V$$

for arbitrary hypergraphs. The coefficients $\left\lfloor \frac{|\{v \in V' : e \in \delta(v)\}|}{p} \right\rfloor$ may now have a value greater than 1.

As in the last step, a Chvátal-Gomory procedure with coefficient $\frac{1}{p}$ for all inequalities of type (SSP) (i) for $v \in V'$ and 0 for all others yields these inequalities.

Step 3. For the third step, observe that for every vertex v in a graph or hypergraph, $\delta(v)$ is a hyperedge clique. To get the odd set inequalities or their generalizations in Steps 1 and 2, the Chvátal-Gomory procedure could be applied to the inequalities of type (SSP) (i), which are clique inequalities. In a graph, $\delta(v)$ is the only type of maximal edge cliques. However, there may be other maximal hyperedge cliques and therefore also other valid clique inequalities for a hypergraph. Applying the previous ideas to also other types of hyperedge cliques for some hyperedge clique set $\mathcal{Q}' \subseteq \mathcal{Q}$ we get the the inequalities

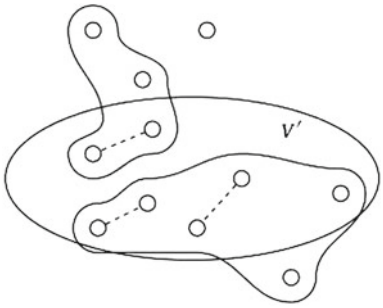


Fig. 2 A packing in a hypergraph $G = (V, E)$ with eleven vertices and a vertex subset V' surrounded by an ellipse, $p = 2$. There can be at most $\lfloor \frac{7}{p} \rfloor = 3$ edges which are subsets of both the seven vertices in V' and some hyperedge in the packing. All other possible edges that would connect two vertices that are contained in some hyperedge in the packing have to have at least one vertex from $V \setminus V'$

$$\sum_{e \in E} \left\lfloor \frac{|\{Q \in \mathcal{Q}' : e \in Q\}|}{p} \right\rfloor x_e \leq \left\lfloor \frac{|\mathcal{Q}'|}{p} \right\rfloor \quad \forall \mathcal{Q}' \subseteq \mathcal{Q},\, p \in \mathbb{N},\, p \geq 2.$$

We remark that for the hypergraph assignment problem (HAP) [2], a partitioning problem on bipartite hypergraphs, for which all inequalities valid for the corresponding set packing relaxation are valid, these inequalities can be facet-defining. In the HAP polytope for a certain "complete bipartite hypergraph with three parts in each of the two vertex sets, all parts having size two", one half of the 30 facet classes (this is modulo symmetry, they contain all together 14049 facets) can be described in this way with $p = 2$.

3 Comparison with General Clique Family Inequalities

Pêcher and Wagler [5] propose a different generalization of odd set cuts of the set packing problem. These inequalities, "general clique family inequalities", have a similar structure (division by some $p \in \mathbb{N}$, rounding, coefficient for a hyperedge variable depends on the number of hyperedge cliques that contain this hyperedge), however, the resulting inequality is different. Also, to the best of our knowledge no combinatorial interpretation was developed for general clique family inequalities so far.

General clique family inequalities are defined as follows. Let $\mathcal{Q}' \subseteq \mathcal{Q}$ be a set of at least three edge cliques for the hypergraph $G = (V, E)$. Choose an integer p with $2 \leq p \leq |\mathcal{Q}'|$, define $R := |\mathcal{Q}'| \mod p$ and choose an integer J with $0 \leq J \leq p - R$. Now define $E_i := \left\{e \in E : |\{Q \in \mathcal{Q}' : e \in Q\}| = i\right\}$ for $i \in \{1, 2, \ldots, |\mathcal{Q}'|\}$ to be the set of hyperedges that are contained in exactly i hyperedge cliques in $\mathcal{Q}'$. The general clique family inequality

Table 1 Coefficient of x_e, $e \in E$ on left hand sides of the inequalities (3) derived in Step 3 and general clique family inequalities (4) depending on the number $i := |\{Q \in \mathcal{Q}' : e \in Q\}|$ of hyperedge cliques in $\mathcal{Q}'$ that contain e

	(3)	(4)
$0 \le i < p - J$	0	0
$p - J \le i < p$	0	$\frac{i-R}{p-R}$
$p \le i \le \lvert\mathcal{Q}'\rvert$	$\left\lfloor \frac{i}{p} \right\rfloor$	1

$$\sum_{i=p}^{|\mathcal{Q}'|}(p-R)\sum_{e\in E_i} x_e + \sum_{j=1}^{J}(p-R-j)\sum_{e\in E_{p-j}} x_e \le b$$

is valid if $b \ge (p-R)\left\lfloor \frac{|\mathcal{Q}'|}{p} \right\rfloor$.

To compare the general clique family inequalities to our inequalities we rewrite Step 3 as

$$\sum_{i=0}^{|\mathcal{Q}'|}\left\lfloor \frac{i}{p} \right\rfloor \sum_{e\in E_i} x_e \le \left\lfloor \frac{|\mathcal{Q}'|}{p} \right\rfloor, \tag{3}$$

and divide both sides of the general clique family inequalities with strongest allowed b by $(p-R)$ to get the valid inequality

$$\sum_{i=p}^{|\mathcal{Q}'|}\sum_{e\in E_i} x_e + \sum_{j=1}^{J}\frac{p-R-j}{p-R}\sum_{e\in E_{p-j}} x_e \le \left\lfloor \frac{|\mathcal{Q}'|}{p} \right\rfloor. \tag{4}$$

Now the right hand sides are equal. The coefficients of x_e, $e \in E$ on the left hand sides are summarized in Table 1 depending on the number $i := |\{Q \in \mathcal{Q}' : e \in Q\}|$ of edge cliques in $\mathcal{Q}'$ that contain e. The table shows that the inequalities concentrate on coefficients for different kinds of hyperedge variables although they employ similar objects. The inequalities derived in this paper have non-zero coefficients only for hyperedges of size $\ge p$. These coefficients may differ depending on the hyperedge size and be > 1, whereas the corresponding coefficients in the general clique family inequalities are all equal to 1. General clique family inequalities, however, have non-zero coefficients for smaller hyperedges.

Thus, the inequality class presented in this paper is different from the general clique family inequalities.

References

1. Borndörfer, R. (1998). Aspects of set packing, partitioning, and covering, PhD thesis, TU Berlin.
2. Borndörfer, R., Heismann, O. (2012). The hypergraph assignment problem. Technical Report 12–14, ZIB.
3. Edmonds, J. (1965). Maximum matching and a polyhedron with 0, 1-vertices. *Journal of Research of the National Bureau of Standards*, *69*, 125–130.
4. Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness (Series of books in the mathematical sciences)*. San Francisco: W.H. Freeman.
5. Pêcher, A., & Wagler, A. (2006). Generalized clique family inequalities for claw-free graphs. *Electronic Notes in Discrete Mathematics*, *25*, 117–121.

New Lower Bounds for the Three-Dimensional Strip Packing Problem

Kirsten Hoffmann

Abstract In this paper, we study the three-dimensional strip packing problem (SPP3) which involves packing a set of non-rotatable boxes into a three-dimensional strip (container) of fixed length and width but unconstrained height. The goal is to pack all of the boxes orthogonal oriented and without overlapping into the container, minimising its resulting height. We present new lower bounds derived from different relaxations of the mathematical formulation of the SPP3. Furthermore, we show dominance relations between different bounds and limit the worst case performance ratio of some bounds.

1 Introduction

In production and transportation, growing scarcity of resources and competition pressure will force companies into efficient management of raw material or storage area. Such optimisation problems can be modelled as bin or strip packing problems.

These problems are NP-hard in the strong sense and, therefore, lower bounds are necessary to limit the optimal solution and to estimate the performance of some heuristic solutions.

For some more details about the mentioned problems, we refer to the typologies of [13, 21]. Lower bounds for the two-dimensional bin packing problem are proposed by [2, 6, 7, 10, 19], whereas the three-dimensional case is discussed in [5, 18]. Surveys concerning lower bounds for the two-dimensional strip packing problem can be found in [1, 8]. See [17] for lower bounds based on geometric considerations and [4] for lower bounds based on relaxations of the original problem. General bounds applicable for several packing problems are given in [14]. Considerably

K. Hoffmann (✉)
Fakultät Wirtschaftswissenschaften, TU Dresden, 01062 Dresden, Germany
e-mail: kirsten.hoffmann@tu-dresden.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_27,

less literature exists for the three-dimensional strip packing problem (also known as container loading problem), especially concerning lower bounds.

Therefore, this article is intended to fill this gap by proposing new lower bounds for SPP3. Therefore, the remainder is organized as follows: In the next section we specify some definitions and simple lower bounds. Lower bounds based on relaxation of the original problem are presented in Sect. 3. The main findings as well as possible directions of future research are summarized in Sect. 4.

2 Problem Description and Simple Bounds

We consider the three-dimensional strip packing problem (SPP3): Given a three-dimensional strip (container) of fixed length L and width W but unconstrained height and a list $L = \{R_1, R_2, \ldots, R_n\}$ of items (boxes) of length $l_i \leq L$, width $w_i \leq W$, and height h_i ($i \in I = \{1, \ldots, n\}$), the goal is to pack the boxes into the container minimising its resulting height. Additional constraints are that the boxes must not overlap and must be packed orthogonally and non-rotatable. All given informations (dimensions of the container and boxes) are associated with the related SPP3-instance.

Definition 1 Let E be an instance of SPP3, $H^*(E)$ the corresponding minimal strip height of the optimal packing, and $b(E)$ the value provided by the lower bound b. The absolute worst-case performance ratio (WCPR) of b is

$$\sup_E \frac{H^*(E)}{b(E)}. \tag{1}$$

Definition 2 Given two different lower bounds b_1 and b_2, b_1 dominates b_2 if and only if $b_1(E) \geq b_2(E)$ for all instances E.

Two simple lower bounds for SPP3 are the continuous lower bound b_1 and the height of the tallest box b_2:

$$b_1 = \left\lceil \frac{1}{LW} \sum_{i \in I} l_i w_i h_i \right\rceil \quad \text{and} \quad b_2 = \max_{i \in I} \{h_i\}. \tag{2}$$

Each bound separately is arbitrarily bad for some instances, but by combining, we get the improved bound

$$b_0 = \max\{b_1, b_2\}. \tag{3}$$

We hazard a guess that the WCPR of b_0 equals 4, but we can proof:

Theorem 1 *The absolute WCPR of* b_0 *is between* 4 *and* 6.

Proof Consider an instance with m boxes (m is a factor of 4) with $l_i = L/2 + \varepsilon$, $w_i = W/2 + \varepsilon$ and $h_i = 1$ for some small $\varepsilon > 0$. Therefore, the minimal container

height equals m and $b_0 = b_1 = m/4 + 1$ for $\varepsilon \to 0$. For sufficient large m the quotient H^*/b_0 is arbitrarily close to 4. On the other hand, $H^* \leq 6b_0$ follows from $H^* \leq 4b_1 + 2b_2$ that is proved in [12]. □

3 Lower Bounds Based on Mathematical Formulations

For SPP3 there are several mathematical models that observe the constraints concerning non-overlapping, orthogonality and non-rotation. See [2, 3, 11, 20] for more details. In practical matters these models can barely be solved optimal. Thus, relaxations are used to approximate the optimal solution.

In optimisation relaxation is a strategy to reduce the difficulty of the original problem in such way that the relaxed problem provides a lower bound (in case of minimisation problems) for the solution of the original problem. For the strip packing problem we can relax the box geometry in one or two dimensions (slice or bar relaxation). Furthermore, we can distinguish between the kind of patterns. For the slice relaxation we can generate feasible two-dimensional patterns or (maybe infeasible) patterns satisfying the knapsack condition. In case of bar relaxation the original problem is reduced to a (contiguous) cutting stock problem.

Exemplary, we observe the slice relaxation along the height H. Either $a^j \in \{0, 1\}^m$, $j \in J$, denotes a feasible two-dimensional pattern or a one-dimensional pattern satisfying the knapsack condition

$$\sum_{i \in I} a_{ij} l_i w_i \leq LW \tag{4}$$

for a slice with length L and width W. x_j represents the quantity of pattern j in the solution, J all possible patterns. The slice relaxation concerning the height is:

$$\sum_{j \in J} x_j \to \min \tag{5}$$

$$\text{s.t.} \sum_{j \in J} a_{ij} x_j = h_i, \quad i \in I, \tag{6}$$

$$x_j \in \mathbb{N}, \quad j \in J. \tag{7}$$

Let z^{b-H} identify the optimal value of (5)–(7) with feasible 2D-patterns, z^{ks-H} with knapsack-patterns.

Theorem 2 *The optimal values z^{b-H} and z^{ks-H} as well as the optimal values of their linear programming (LP) relaxation of (5)–(7) denoted by z_{LP}^{b-H} and z_{LP}^{ks-H} dominate b_0.*

Proof We have to show that all optimal values of the relaxations dominate both b_1 and b_2. The transformation of constraint (6) yields to

$$\sum_{j\in J} l_i w_i a_{ij} x_j = l_i w_i h_i, \quad i\in I$$

$$\sum_{j\in J} x_j \sum_{i\in I} a_{ij} l_i w_i = \sum_{i\in I} l_i w_i h_i.$$

For both kind of patterns the inequality $\sum_{i\in I} a_{ij} l_i w_i \le LW$ stays true. Therefore, we have $z^{b-H} \ge b_1$ and $z^{ks-H} \ge b_1$. Furthermore, with

$$\sum_{j\in J} x_j \ge \sum_{j\in J} a_{ij} x_j = h_i \ge h_{\max}, \quad i\in I$$

we can show $z^{b-H} \ge b_2$ and $z^{ks-H} \ge b_2$. Because these relations hold for both $x_j \in \mathbb{N}$ and $x_j \in \mathbb{R}_+$ (LP relaxation) we complete the proof. □

Theorem 3 *For the absolute WCPR we have*

$$2 \le \sup_E \frac{H^*}{z^{b-H}} \le 6 \quad \textit{and} \quad 3 \le \sup_E \frac{H^*}{z^{ks-H}} \le 6. \tag{8}$$

Proof Because of theorem 1 and 2 we have

$$H^* \le 6b_0 \le 6z^{b-H} \quad \text{and} \quad H^* \le 6b_0 \le 6z^{ks-H}.$$

Consider the instance with $k+1$ boxes, $l_i = 1$, $w_i = h_i = k$ and $L = W = k$. The ratio $H^*/z^{b-H} = 2k/(k+1)$ is arbitrarily close to 2 for sufficient large k. Consider now the instance with m(m is a factor of 3) boxes, $l_i = L/2 + \varepsilon$, $w_i = W/2 + \varepsilon$ for small $\varepsilon > 0$ and $h_i = 1$. We have $H^*/z^{ks-H} = m/(m/3) = 3$ for sufficient small ε. □

For further relaxation we can group the boxes by the dimensions of their base area. Let $\overline{m}$ be the number of different base areas $\overline{l}_n \times \overline{w}_n$ and $I_n = \{i \in I : l_i = \overline{l}_n, w_i = \overline{w}_n\}$ for $n \in \overline{I} = \{1, \ldots, \overline{m}\}$. $a^j \in \mathbb{Z}_+^{\overline{m}}$ either denotes a feasible two-dimensional pattern or a one-dimensional pattern satisfying the knapsack condition

$$\sum_{n\in\overline{I}} a_{nj} l_n w_n \le LW \tag{9}$$

for a slice with length L and width W. The slice relaxation concerning the height with grouping the boxes is:

$$\sum_{j\in J} x_j \to \min \tag{10}$$

$$\text{s.t.} \sum_{j\in J} a_{nj}x_j = \sum_{i\in I_n} h_i, \quad n \in \bar{I}, \tag{11}$$

$$x_j \in \mathbb{N}, \quad j \in J, \tag{12}$$

where x_j represents the quantity of pattern j in the solution, J all potential patterns. Let z^{g-H} and z^{gks-H} denote the optimal value of (10)–(12) with feasible 2D-patterns and knapsack-patterns respectively.

The relaxation of the original problem has been pushed so far that these bounds are arbitrarily bad, i. e., the absolute worst-case performance cannot be determined. However, we can make a proposition about the asymptotic worst-case behaviour.

Theorem 4 *Let $\overline{m}$ be the number of different base areas of boxes L and $h_{\max} = \max_{i\in I}\{h_i\}$ the height of the tallest box. For the minimal strip height we have*

$$H^* \le z^{g-H} + \overline{m}h_{\max}. \tag{13}$$

Proof Note that in any basic (extreme point) solution to (10)–(12) the number of non-zero coordinates of the solution is at most the number of constraints, excluded the non-negativity constraints. Let $(x_1, \ldots, x_{|J|})$ denote the solution of (10)–(12), thus, we have m' non-zero coordinates $x_1, \ldots, x_{m'}$ with $m' \le \overline{m}$. The algorithm, that generates a feasible three-dimensional strip packing of height $z^{g-H} + \overline{m}h_{\max}$, works as follows: Starting at $j = 1$, create a level of height $x_j + h_{\max}$ for each j with $x_j > 0$. For each n with $a_{nj} \neq 0$ draw a_{nj} columns with length $\bar{l}_n$ and width $\overline{w}_n$ covering the total height. Afterwards fill all columns with length $\bar{l}_n$ and width $\overline{w}_n$ with boxes of I_n in a greedy manner.

Suppose that all boxes fit into the container. The proof is by contradiction. Assume a box s with length $\bar{l}_n$ and width $\overline{w}_n$ does not fit in any column with such dimensions. The height of s is at most $h_{\max}$, whereas the height of the columns equals $x_j + h_{\max}$ for some x_j. Since box s does not fit, all columns must be filled up to more than x_j. The cumulative height of all boxes already placed in these columns with length $\bar{l}_n$ and width $\overline{w}_n$ is more than $\sum_{j\in J} a_{nj}x_j = d_n$, which leads to a contradiction.

The proposed algorithm yields a feasible strip packing for all boxes. The total height equals $(x_1+h_{\max})+\cdots+(x_{m'}+h_{\max}) = z^{g-H}+m'h_{\max} \le z^{g-H}+\overline{m}h_{\max}$. □

Notice that if $h_{\max}$ and $\overline{m}$ are bounded, the WCPR converges asymptotically to one. A related theorem for the two-dimensional case can be found in [16].

Theorem 5 *Consider bins of size $L \times W$ and a list of rectangular items with $\max_{i\in I} l_i \le L$ and $\max_{i\in I} w_i \le W$. Then all items can be packed into at most three bins, if $\sum_{i\in I} l_i w_i \le LW$.*

Proof See [9, 15]. □

Theorem 6 *Let $\overline{m}$ be the number of different base areas of boxes $i \in I$ and $h_{\max} = \max_{i \in I}\{h_i\}$ the height of the tallest box. For the minimal strip height we have*

$$H^* \leq 3(z^{gks-H} + \overline{m}h_{\max}). \tag{14}$$

Proof Theorem 5 proved that the base areas of a subset K of boxes that satisfy the knapsack condition $\sum_{i \in K} l_i w_i \leq LW$, can be packed into at most three bins of size $L \times W$. Further steps are analogous to the proof of theorem 4 with the exception that for each j with $x_j > 0$ now three levels of height $x_j + h_{\max}$ are created. □

4 Conclusion

The three-dimensional strip packing problem considered in this paper has several important applications to practical problems, e. g., the container loading problem. It is NP-hard and, therefore, optimal solving is fairly complicated or even impossible. We propose new lower bounds for this problem. Furthermore, this paper provides results on the worst-case performance ratio (absolute or asymptotic) of the proposed bounds. Further research should ask for more and improved lower bounds and perform computational tests.

References

1. Alvarez-Valdés, R., Parreño, F., & Tamarit, J. M. (2009). A branch and bound algorithm for the strip packing problem. *OR Spectrum*, *31*(2), 431–459.
2. Beasley, J. E. (1985). Bounds for two-dimensional cutting. *The Journal of the Operational Research Society*, *36*(1), 71–74.
3. Belov, G., Kartak, V. M., Rohling, H., & Scheithauer, G. (2009). One-dimensional relaxations and LP bounds for orthogonal packing. *International Transactions in Operational Research*, *16*(6), 745–766.
4. Belov G., Scheithauer G. (2011) *Gaps between optimal values of some packing and scheduling problems*. Technische Universität Dresden.
5. Boschetti, M. A. (2004). New lower bounds for the three-dimensional finite bin packing problem. *Discrete Applied Mathematics*, *140*(1), 241–258.
6. Boschetti, M. A., & Mingozzi, A. (2003). The two-dimensional finite bin packing problem. Part I: New lower bounds for the oriented case. *4OR: A Quarterly Journal of Operations Research*, *1*(1), 27–42.
7. Boschetti, M. A., & Mingozzi, A. (2003). The two-dimensional finite bin packing problem. Part II: New lower and upper bounds. *4OR: A Quarterly Journal of Operations Research*, *1*(2), 135–147.
8. Boschetti, M. A., & Montaletti, L. (2010). An exact algorithm for the two-dimensional strip-packing problem. *Operations Research*, *58*(6), 1774–1791.
9. Buchwald, T., Hoffmann, K., & Scheithauer, G. (2013). *Relations between capacity utilization and minimal bin number*. Technical Report, TU Dresden.

10. Carlier, J., Clautiaux, F., & Moukrim, A. (2007). New reduction procedures and lower bounds for the two-dimensional bin packing problem with fixed orientation. *Computers and Operations Research*, *34*(8), 2223–2250.
11. Chen, C., Lee, S., & Shen, Q. (1995). An analytical model for the container loading problem. *European Journal Of Operational Research*, *80*(1), 68–76.
12. Diedrich, F., Harren, R., Jansen, K., Thöle, R., & Thomas, H. (2008). Approximation algorithms for 3D orthogonal Knapsack. *Journal of Computer Science and Technology*, *23*(5), 749–762.
13. Dyckhoff, H. (1990). A typology of cutting and packing problems. *European Journal of Operational Research*, *44*(2), 145–159.
14. Fekete, S. P., & Schepers, J. (2004). A general framework for bounds for higher-dimensional orthogonal packing problems. *Mathematical Methods of Operations Research*, *60*(2), 311–329.
15. Hoffmann, K. (2012). Das Streifenpackproblem: Untere Schranken und ihre Güte. Masters thesis, TU Dresden.
16. Kenyon, C., & Rémila, E. (2000). A near-optimal solution to a two-dimensional cutting stock problem. *Mathematics of Operations Research*, *25*(4), 645–656.
17. Martello, S., Monaci, M., & Vigo, D. (2003). An exact approach to the strip-packing problem. *INFORMS Journal on Computing*, *15*(3), 310–319.
18. Martello, S., Pisinger, D., & Vigo, D. (2000). The three-dimensional bin packing problem. *Operations Research*, *48*(2), 256–267.
19. Martello, S., & Toth, P. (1990). Lower bounds and reduction procedures for the bin packing problem. *Discrete Applied Mathematics*, *28*(1), 59–70.
20. Padberg, M. (2000). Packing small boxes into a big box. *Mathematical Methods of Operations Research*, *52*(1), 1–21.
21. Wäscher, G., Haußner, H., & Schumann, H. (2007). An improved typology of cutting and packing problems. *European Journal of Operational Research*, *183*(3), 1109–1130.

A New Method for Parameter Estimation of the GNL Model Using Real-Coded GA

Yasuhiro Iida, Kei Takahashi and Takahiro Ohno

Abstract In this paper, a new parameter estimation method is proposed for the generalized nested logit (GNL) model using real-coded genetic algorithms (GA). We propose a method to recalculate and verify whether the offsprings violate constraints. In addition, we improve the selection and mutation operators in order to find the higher log likelihood. In the numerical experiments, the log likelihood of our method is compared to that obtained by the Quasi-Newton method and the normal real-coded GA, which use SPX and JGG, and not the mutation operator, with the actual point of sales data. Thus, we prove that our method finds a higher log likelihood than conventional methods.

1 Introduction

The generalized nested logit (GNL) model is a discrete choice model used in transportation, route choice, and brand choice to represent a selection of one among a set of mutually exclusive alternatives [3]. When we perform the parameter estimation of the GNL model, the log likelihood maximization is usually used. Furthermore,

Y. Iida (✉)
Department of Business Design and Management, Graduate School of Waseda University,
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: y-iida@ruri.waseda.jp

K. Takahashi
School of Statistical Thinking, The Institute of Statistical Mathematics,
10-3 Midoricho, Tachikawa-shi, Tokyo 190-8562, Japan
e-mail: k-taka@ism.ac.jp

T. Ohno
Department of Industrial and 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

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_28,

in the log likelihood maximization algorithm, the Quasi-Newton method or steepest descent method is widely used. The log likelihood function of the GNL model is multimodal. Therefore, when we use those methods, the result of the parameter estimation depends on its initial settings and we may find the local optima [4]. In this study, we improve the real-coded GA for parameter estimations of the GNL model in order to find the quasi-optima in each estimation for different initial settings.

2 The GNL Model

The choice probability of the GNL model for which consumer i chooses alternative n assigned to nest m can be expressed as

$$Q_n^i = \sum_m P_m^i P_{n|m}^i, \tag{1}$$

$$P_m^i = \frac{\left(\sum_{n' \in N_m} \left(\alpha_{n'm} \exp(V_{n'}^i)\right)^{\frac{1}{\mu_m}}\right)^{\mu_m}}{\sum_m \left(\sum_{n' \in N_m} \left(\alpha_{n'm} \exp(V_{n'}^i)\right)^{\frac{1}{\mu_m}}\right)^{\mu_m}}, \tag{2}$$

$$P_{n|m}^i = \frac{\left(\alpha_{nm} \exp(V_n^i)\right)^{\frac{1}{\mu_m}}}{\sum_{n' \in N_m} \left(\alpha_{n'm} \exp(V_{n'}^i)\right)^{\frac{1}{\mu_m}}}, \tag{3}$$

where P_m^i denotes the choice probability of nest m, $P_{n|m}^i$ is the choice probability of alternative n if nest m is selected, V_n^i is the utility for each alternative n, N_m is the set of all alternatives included in nest m, μ_m is the logsum parameter for nest m and α_{nm} is the allocation parameter that characterizes the portion of alternative n assigned to nest m. Furthermore, in order to be consistent with random utility maximization μ_m and α_{nm} must satisfy the following conditions $0 < \mu_m \leq 1$, $\alpha_{nm} \geq 0$, and

$$\sum_m \alpha_{nm} = 1. \tag{4}$$

The log likelihood function of the GNL model can be expressed as

$$\ln L = \sum_i \sum_m \sum_n \mathbf{1}_{mn}^i \left(\ln P_{n|m}^i + \ln P_m^i\right), \tag{5}$$

where $\mathbf{1}_{mn}^i$ is a 0-1 variable denoting whether the alternative n assigned to nest m is chosen (0) or not chosen (1).

3 Real-Coded GA

In this section, we introduce real-coded GA. A real-coded GA has three operators, (1) crossover operator, (2) selection operator and (3) mutation operator.

3.1 Crossover Operator

We apply simplex crossover (SPX) to the parameter estimation of the GNL model. Conventional studies show that SPX optimizes the various test functions efficiently and its performance is independent of any linear coordinate transformation [2]. In SPX, the offsprings are generated by the following five-step procedure:

1. Select $n+1$ parents $x^0, \ldots, x^n$ randomly from the current population.
2. Let the gravity point of the parents be g.
3. Generate random number r_k as $r_k = (u(0,1))^{\frac{1}{k+1}}$, where $u(0,1)$ is a uniform random number.
4. Calculate p^k and c^k as, respectively,

$$p^k = g + \alpha(x^k - g), \qquad \text{for } k = 0, 1, \ldots, n,$$

$$c^k = \begin{cases} 0 & \text{for } k = 0, \\ r_{k-1}(p^{k-1} - p^k + c^{k-1}) & \text{for } k = 1, \ldots, n. \end{cases}$$

5. Generate an offspring x^c as $x^c = p^n + c^n$.

However, when we apply SPX to the parameter estimation of the GNL model that has constraints in allocation and logsum parameters (Eq. 4), there is a possibility of generating offsprings that may violate the constraints. Therefore, we propose a method to recalculate and verify whether the offsprings satisfy constraints. The procedure is shown in Fig. 1.

3.2 Selection Operator

We propose reJGG based on just generation gap (JGG). In reJGG, parents and offsprings are selected by the following three-step procedure:

1. Select n_p parents randomly from the current population.
2. Apply the crossover operator and generate n_c offsprings.
3. Replace the parents that were used with the crossover operator with the top n_p parents and offsprings.

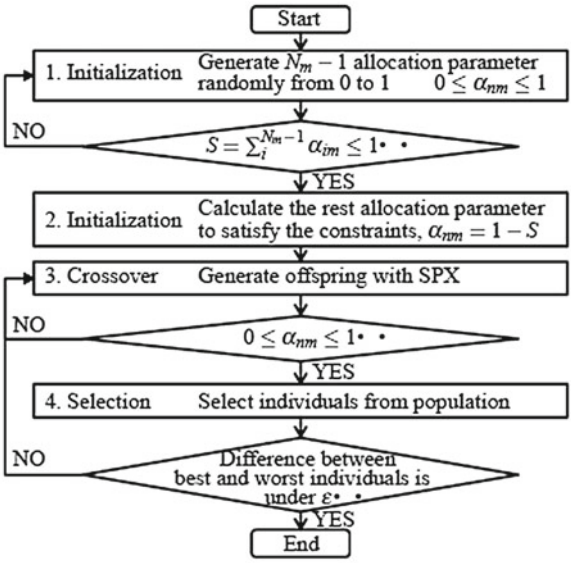


Fig. 1 The procedure for recalculating if offsprings do not satisfy constraints

3.3 Mutation Operator

We propose the use of mutation operator, which is not generally used in real-coded GA, to maintain diversity and not for the purpose of finding the local optima. The GNL model involves allocation parameters that have constraints in each alternative, $\sum_m \alpha_{nm} = 1$. When we perform parameter estimation of the GNL model, some allocation parameters converge quickly, whereas others do not. Therefore, some allocation parameters that almost converge might not be global optima. In the first step, we apply a uniform mutation to the parameter estimation process and prove the significance of using a mutation operator. In the second step, we improve the uniform mutation.

1. Generate a new real-valued number randomly with mutation rate p
2. Generate a new real-valued number randomly with mutation rate p if the ratio of the area made by parents to that made by constraints falls below the area rate p_a.

4 Numerical Experiment

We test the performance of our method and compare it with that of the Quasi-Newton method and normal real-coded GA, which use SPX and JGG, and not the mutation operator. In this study, we use the scanner-panel data in a supermarket. The data

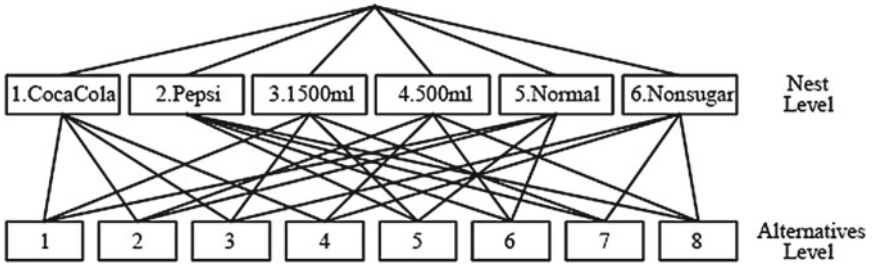


Fig. 2 Structure of GNL model. The upper level indicate the nest that groups the alternatives in the same category and the lower level is the alternatives assigned to each nest

Table 1 Parameters for real-coded GA

Parameter	Value
Number of parameter	34
Initial population	340
Iteration number of crossover operator	340
Mutation rate	0.1
Number of trials	20

includes 5,269 purchases and the GNL model structure is shown in Fig. 2. In the GNL model, the utility for each alternative n can be represented as

$$V_n^i = \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \beta_4 X_{4n} + \beta_5 X_{5n}, \tag{6}$$

where X_{1n} denotes the price of alternative n, X_{2n} is a 0-1 variable that denotes whether the consumer i chooses CocaCola, X_{3n} is the volume of alternative n, X_{4n} is a 0-1 variable that denotes whether the alternative n's content is nonsugar, X_{5n} is a 0-1 variable that denotes whether the consumer i purchases the same alternative previously and β_i is utility parameter. Table 1 shows the log likelihood and calculating time of the Quasi-Newton method and that of the real-coded GA. We can not find the optima using the Quasi-Newton method. Owing to the application of real-coded GA, the log likelihood is increased. The result indicates superiority in performance of real-coded GA. Next, we compare the combination of JGG or reJGG along with the use of mutation operator and without its use with each real-coded GA. Due to the application of reJGG, the log likelihood is increased and the calculating time is reduced. This is because reJGG does not discard the parent that has good utility. Therefore, it is easy to generate offsprings that speed up the convergence and have a better value. By applying uniform mutation, the log likelihood and calculating time are increased. This is because the mutation operator searches a wider area. This makes the convergence slow and it is easy to find offsprings that have a better value.

Table 2 Comparison the Quasi-Newton method with the real-coded GA. In the Quasi-Newton method, we test 40 trials from different initial settings

Method	Log likelihood	Time (min)
Quasi-Newton method	−5820.54	–
SPX + JGG	−5803.03	47.29
SPX + reJGG	−5801.00	42.14
SPX + JGG + Mutation	−5802.45	47.08
SPX + reJGG + Mutation	−5800.90	44.95

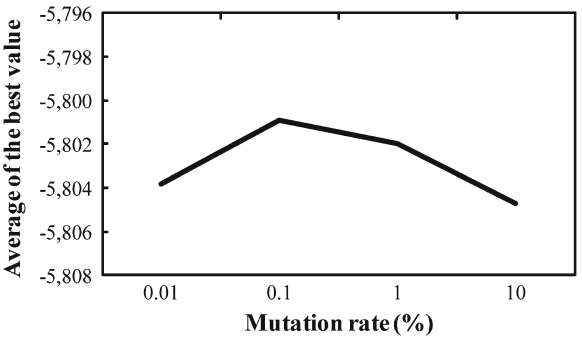


Fig. 3 The average of the best values obtained when mutation rate is changed

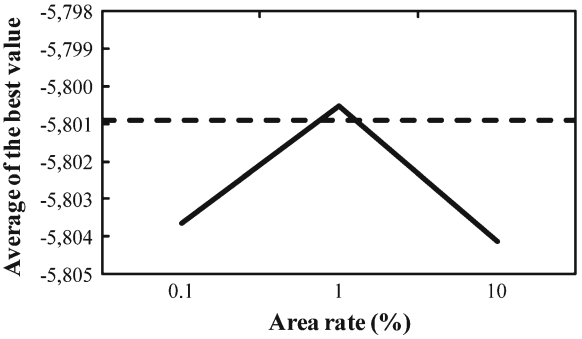


Fig. 4 The average of the best values obtained when mutation operator is changed. The *dash line* indicates the result of Step 1 when mutation rate $p = 0.1$ and the *continuous line* indicates the result of Step 2

We observed that our method, which uses SPX, reJGG, and uniform mutation, finds a higher log likelihood compared to conventional methods. Therefore, we continued with this method and changed some parameters. First, we changed the mutation rate p. Figure 3 shows the result. As seen in Fig. 3, we obtained the best value at $p = 0.1$. If we use a very small mutation rate, we can find the lower log likelihood because a very small mutation rate is similar to no mutation operator. Furthermore,

if we use a very large mutation rate, we can find the lower log likelihood because a very large mutation rate is similar to random searching. Next, we implement the second step. Figure 4 shows the result. As seen in 4, we obtained the best value for $p_a = 1$. We can also find the local optima if we use a very small or large area rate.

5 Conclusion

We introduced a parameter estimation method for the GNL model using real-coded GA. To improve the accuracy of the estimation, we proposed two algorithms, reJGG and mutation operator. Using our method, the log likelihood is increased and the calculating time is reduced. Further investigations into the calculating time and other methods are needed to obtain more precise outcomes.

References

1. Fish, K. E., Johnson, J. D., Dorsey, R. E., & Blodgett, J. G. (2004). Using an artificial neural network trained with a genetic algorithm to model brand share. *Journal of Business Research*, *57*, 79–85.
2. Higuchi, T., Tsutsumi, S., & Yamamura, M. (2000) Theoretical analysis of simplex crossover for real-coded genetic algorithms. In *Parallel problem solving from nature* (pp. 365–374).
3. Wen, C. H., & Koppelman, F. S. (2001). The generalized nested logit model. *Transportation Research Part B*, *63*(2), 627–641.
4. Zhuang, X. Y., Fukuda, D., & Yai, T. (2007). Analyzing inter-regional travel mode choice behavior with multi nested generalized extreme value model. *Journal of the Eastern Asia Society for Transportation Studies*, *7*, 686–699.

Inventory Control with Supply Backordering

Marko Jakšič

Abstract We study the inventory control problem of a retailer working under stochastic demand and stochastic limited supply. The unfulfilled part of the retailer's order is backordered at the supplier and that the retailer has a right to cancel the replenishment of the backordered supply, if desired. We show the optimality of the a base-stock type type policy and derive the threshold inventory position over which it is optimal to cancel the replenishment of the backordered supply. We carry out a numerical analysis to quantify the benefits of supply backordering and the value of the cancelation option.

1 Introduction

The importance of time as a competitive weapon in supply chains has been recognized for some time. Suppliers venture into lead time projects to improve the ability to meet the demands of customers for shorter lead times. However, such undertaking often results in, at least in a short term, worse supply performance, characterized mainly by delayed and/or partial replenishment.

In this paper we study the inventory control problem of a retailer working under stochastic demand from the market, where he tries to satisfy the demand by making orders with a supplier. The supply capacity available to the retailer is assumed to be limited and stochastic as a result of a supplier's changing capacity and capacity allocation policy. The order placed by the retailer might therefore not be delivered in full, depending on the currently available capacity. The novel feature of our model is that the unfulfilled part of the retailer's order is backordered at the supplier. We assume that the replenishment of the backordered supply is certain, meaning that it

M. Jakšič (✉)
Department of Management and Organization, Faculty of Economics, University of Ljubljana, Kardeljeva ploščad 17, Ljubljana, Slovenia
e-mail: marko.jaksic@ef.uni-lj.si

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_29,

is delivered in full in the following period (together with replenishment of the next period's regular order). As the supply backorder is a result of the supplier's inadequate supply service, this gives the retailer an option (a moral right) to cancel the replenishment of the backordered supply if necessary. Therefore, in each period the retailer has to make two decisions. Apart from the regular ordering decision to the supplier, he needs to decide whether he wants the supplier to replenish the backordered supply or not. We denote the case where the backordered supply is always replenished as the *Full backordering* (*FB*) policy, and compare it to the *No backordering* (*NB*) policy, where there is no supply backordering, to establish the value of supply backordering. We are also interested in whether full supply backordering can be counterproductive in specific situations. Therefore, we quantify the effect of the retailer's ability to cancel the supply backorder at the supplier (through implementing a *Cancelation option* (*CO*) policy) on the reduction of inventory costs.

The way we model the supply availability is inline with the work of [1–3], where the random supply/production capacity determines a random upper bound on the supply availability in each period. A general assumption in capacitated inventory models is that the part of the order above the available supply capacity in a certain period is lost to the customer. We believe this might not hold in several situations observed in practice.

While demand backordering a strategy extensively used on the demand side, supply backorders have not been considered in the literature. For instance, in [4], supplier gives priority to satisfying the backordered part of the demand from previous periods. Again considering the demand side, a stream of research deals with the problem of demand or order cancelation. In [5] they assume a constant fraction of customers are canceling their backorders. Therefore they do not consider the cancelation of backorders as a decision variable, but as a preset system parameter, which effectively reduces the demand the supplier is facing. Such treatment is essentially similar to so called *partial backordering* that is considered in the inventory literature for cases where part of the demand is backordered while the remainder is lost, as in [6, 7]. In this paper we include the option to cancel backorders on the supply side as an integral part of the optimal ordering decision policy.

In Sect. 2 we present our dynamic programming model incorporating supply backordering. The structure of the optimal policy is given in Sect. 3, and the results of a numerical study are presented in Sect. 4. Finally we summarize our findings in Sect. 5.

2 Model Formulation

In this section, we present the dynamic programming model to formulate the problem under consideration. The model assumes a periodic-review inventory control system with non-stationary stochastic demand and limited non-stationary stochastic supply with a zero supply lead time. The supply capacity is assumed to be exogenous to the retailer and the exact capacity realization is only revealed upon replenishment.

Unused capacity in a certain period is assumed to be lost. In the case when currently available supply capacity is insufficient to cover the whole order, a retailer has an option to backorder the unfilled part of the supply at the supplier and it will be delivered to him together with the replenishment of the next order.

We assume the following sequence of events:

(1) At the start of period t, the decision maker reviews the inventory position x_t, and the ordering decision is made, which is composed of two decisions: (1) supply backorder $\beta_{t-1}b_{t-1}$ from the previous period, where the backorder parameter β attains a value of 1 when the backordered supply b_{t-1} is to be replenished, and 0 if it is canceled, and (2) regular order z_t.
(2) The previous period's supply backorder $\beta_{t-1}b_{t-1}$ and the current period's regular order z_t are replenished.
(3) At the end of the period the decision maker observes the previously backordered demand and the current period's demand d_t, and tries to satisfy it from the available inventory $y_t - b_t$. Unsatisfied demand is backordered, and inventory holding and backorder costs are incurred based on the end-of-period inventory position, $x_{t+1} = y_t - b_t - d_t$.

The system's costs consist of inventory holding c_h and backorder c_b costs charged on end-of-period on-hand inventory. The expected single-period cost charged at the end of period t is expressed as $C_t(y_t, z_t) = \alpha E_{Q_t,D_t}\tilde{C}_t(y_t - b_t - D_t)$, where $\tilde{C}(x) = c_h \int_0^x (x - D_t)g_t(D_t)\mathrm{d}D_t + c_b \int_x^\infty (D_t - x)g_t(D_t)\mathrm{d}D_t$ is the regular loss function and $g_t(D_t)$ is a probability density function of demand.

The dynamic programming formulation minimizing the relevant inventory costs over finite planning horizon T from time t onward and starting in the initial state (x_t, b_{t-1}) characterized by the inventory position before the decision making x_t and the backordered supply b_{t-1}, can be written as:

$$f_t(x_t, b_{t-1}) = \min_{\beta_{t-1}=\{0,1\}, z_t \geq 0} \left[C_t(y_t, z_t) + \alpha E_{Q_t,D_t} f_{t+1}(y_t - b_t - D_t, b_t)\right], \text{ for } 1 \leq t \leq T, \quad (1)$$

where the ending condition is defined as $f_{T+1}(\cdot) \equiv 0$.

3 Structure of the Optimal Policy

In this section, we focus on the optimal policy characterization of the inventory system that permits cancelation of the replenishment of the backordered supply. We define the cost function J_t as $J_t(y_t, z_t) = C_t(y_t, z_t) + \alpha E_{Q_t,D_t} f_{t+1}(y_t - b_t - D_t, b_t)$.

In Part 1 of Theorem 1, we show that the inventory policy which minimizes (1) under the *FB* policy, is a base-stock policy characterized by the optimal inventory position after ordering $\hat{y}_t$. Finding the optimal order size $\hat{z}_t$ requires searching for the global minimum of the auxiliary cost function $J_t(y_t, z_t)$, which exhibits a quasiconvex shape and thus has a unique minimum. The quasiconvexity is preserved

through t as the underlying single-period cost function $C(y_t, z_t)$ is also quasiconvex in z_t and function $f_{t+1}(x_{t+1}, b_t)$ for $b_t = 0$ is convex in x_t. The latter holds due to the fact that the first partial derivative of J_t with regard to z_t is independent of b_t. This means that in its general form the problem equals the stochastic capacitated inventory problem studied by [1], where they show the optimality of the base-stock policy.

Theorem 1 *Let $\hat{y}_t$ be the smallest minimizer of $J_t(y_t, z_t)$ and the starting state is (x_t, b_{t-1}):*

1. *The optimal FB policy is a base-stock policy with the optimal base-stock level $\hat{y}_t$ and the optimal order size is $\hat{z}_t = [\hat{y}_t - x_t - b_{t-1}]^+$.*
2. *The threshold inventory position $\bar{y}_t = x_t + \bar{b}_{t-1} \geq \hat{y}_t$, where $\bar{b}_{t-1}$ is a threshold supply backorder level, is a solution to $J_t(x_t, \bar{b}_{t-1}, 0) = J_t(x_t, 0, \hat{z}_t)$, where $\hat{z}_t = \hat{y}_t - x_t$ is the optimal order size.*
3. *Under the optimal CO policy, $y_t(x_t, b_{t-1})$ is given by:*

$$y_t(x_t, b_{t-1}) = \begin{cases} x_t, & \hat{y}_t \leq x_t, & \beta_{t-1} = 0, z_t = 0, \\ x_t + z_t = \hat{y}_t, & x_t < \hat{y}_t < \bar{y}_t \leq x_t + b_{t-1}, & \beta_{t-1} = 0, z_t > 0, \\ x_t + b_{t-1}, & x_t < \hat{y}_t \leq x_t + b_{t-1} < \bar{y}_t, & \beta_{t-1} = 1, z_t = 0, \\ x_t + b_{t-1} + z_t = \hat{y}_t, & x_t + b_{t-1} < \hat{y}_t, & \beta_{t-1} = 1, z_t > 0, \end{cases} \tag{2}$$

4. *The optimal CO policy is a base-stock policy with the optimal base-stock level $\hat{y}_t$, which is equal to the optimal base-stock level of the FB policy, for any t.*

We move to the analysis of the supply backorder cancelation option. In Part 2, we define a threshold inventory position $\bar{y}_t$ above which it is optimal to cancel the replenishment of the supply backorder, as it represents the point at which the costs of either solely replenishing the supply backorder and only placing an optimal regular order up to the optimal base-stock level are equal. Observe that placing a regular order generally results in the inventory position after replenishment $\hat{y}_t - b_t$, which is below the base-stock level due to potential capacity unavailability. On the other hand, replenishing the backordered supply overshoots the base-stock level. As Part 3 suggests, the decision maker should replenish the backordered supply if it is below the threshold size $b_{t-1} \leq \bar{b}_{t-1}$, and in this case no regular order is placed ($\beta_{t-1} = 1, z_t = 0$). When $b_{t-1} > \bar{b}_{t-1}$ it is optimal to cancel the replenishment of the backordered supply and place a regular order up to the optimal base-stock level instead ($\beta_{t-1} = 0, z_t > 0$). In Part 4, we show that the optimal base-stock levels are the same for the *CO* policy and the *FB* policy, which is a consequence of the fact that in both cases the optimal base-stock level is independent of the backordered supply.

4 The Value of Supply Backordering and Cancelation Option

To evaluate the benefits of supply backorder replenishment and the value of the option to cancel the backordered supply, we carried out a numerical analysis. Calculations were done by solving the dynamic programming formulation given in (1),

Table 1 The value of supply backordering and the cancelation option ($T = 12, \alpha = 0.99$, $c_b/c_h = 20$)

$Util$	CV_Q	CV_D	$\%V_{FB}$				$\%V_{CO}$			
			0.00	0.14	0.37	0.61	0.00	0.14	0.37	0.61
∞			100.0	97.1	93.7	90.9	–	0.0	19.2	40.6
2	0.00		100.0	94.8	89.6	87.2	–	7.1	44.4	47.3
2	0.14		96.9	93.4	89.1	87.0	0.0	8.6	39.9	45.0
2	0.37		91.6	90.0	87.5	86.0	0.0	4.6	25.1	35.5
2	0.61		87.4	86.9	85.6	84.7	0.0	1.5	13.8	24.8
1	0.00		–	78.7	80.4	80.6	–	22.9	19.6	23.4
1	0.14		75.7	70.8	77.8	79.3	0.0	3.0	10.8	19.1
1	0.37		71.5	71.0	72.9	75.4	0.0	0.4	2.7	8.3
1	0.61		70.4	71.5	71.7	72.9	0.0	0.1	1.3	3.9
0.67	0.00		–	–	44.9	69.1	–	–	2.0	12.7
0.67	0.14		–	–	34.4	62.4	–	–	0.0	5.8
0.67	0.37		34.3	37.7	46.9	55.9	0.0	0.0	0.0	1.9
0.67	0.61		53.3	54.8	55.7	58.2	0.0	0.0	0.2	0.8

and are presented in Table 1. The value of supply backordering is assessed based on a comparison between the *NB* policy and the *FB* policy ($\%V_{FB}$), while the value of cancelation option is quantified based on the comparison of the *CO* policy and the *FB* policy ($\%V_{CO}$). The relative value $\%V$ is defined as the difference in cost of the policies under consideration relative to the costs of the infinite capacity scenario.

The relative value of supply backordering $\%V_{FB}$ changes considerably over the set of experiments, ranging from scenarios for some of the low utilization experiments denoted with "–", where the three strategies have the same costs, to practically 100 % for high utilization. Due to supply capacity shortages the *NB* policy is unable to cope with the demand, which results in high cost mainly attributed to a high share of backordered demand. The replenishment of supply backorders effectively decreases the system's utilization through the full, albeit postponed, replenishment of orders.

While the value of supply backordering exhibits monotonic behavior with the change in the system's utilization, this is not the case when we consider the effect of demand and/or capacity uncertainty. When the utilization is high ($Util = 2$), $\%V_{FB}$ decreases with the increase in demand uncertainty. In this case, the potentially high supply backorder is fully replenished through *FB* policy, which is not optimal if a low demand period just occurred. For lower utilizations, $\%V_{FB}$ generally increases with CV_D. Here, more stockouts are the result of the target inventory level being insufficient to cover the unusually high demand, and not due to the capacity shortage. The supply backorders are smaller and it is therefore less likely that the replenishment of the supply backorder will be counterproductive in the low demand periods.

The higher the system utilization the higher the value of the cancelation option, where $\%V_{CO}$ for the two high utilization scenarios reaches up to 50 %. As the period-to-period optimal base-stock levels need to be high enough to cover potential future supply shortages, they are sensitive to changes in future period-to-period utilization.

With the flexibility that the *CO* policy offers, it becomes more likely that the optimal base-stock levels will be attained. As low demand periods are more likely to happen when demand uncertainty is high (and these may lead to excessive inventory levels), the costs can be lowered through exercising the cancelation option, however only if the capacity uncertainty is low enough to guarantee a reasonably reliable replenishment of the regular order.

5 Conclusions

In this paper we establish the optimal inventory control policies for a finite horizon stochastic capacitated inventory system in which the unfilled part of an order is backordered at the supplier and delivered in full in the following period; a concept we denote as supply backordering. For both policies, the *Full backordering* and the *Cancelation option*, we show that the structure of the optimal inventory policy is a base-stock policy where the base-stock levels are equal in both cases. We characterize the threshold inventory position above which it is optimal to cancel the replenishment of the backordered supply and place a new order instead. We show that the relative cost savings achieved through supply backordering can be substantial already at moderate system utilization. However these also depend on demand and capacity variability in a complex non-monotonic manner, which requires the decision maker to consider them in an integrated manner. We also establish the following conditions in which exercising the cancelation option is optimal: high system utilization, high demand uncertainty and low capacity uncertainty.

References

1. Ciarallo, F. W., Akella, R., & Morton, T. E. (1994). A periodic review, production planning model with uncertain capacity and uncertain demand—Optimality of extended myopic policies. *Management Science*, *40*, 320–332.
2. Khang, D. B., & Fujiwara, O. (2000). Optimality of myopic ordering policies for inventory model with stochastic supply. *Operations Research*, *48*, 181–184.
3. Iida, T. (2002). A non-stationary periodic review production-inventory model with uncertain production capacity and uncertain demand. *European Journal of Operational Research*, *140*, 670–683.
4. Sox, C. R., Thomas, L. J., & McClain, J. O. (1997). Coordinating production and inventory to improve service. *Management Science*, *43*, 1189–1197.
5. You, P. S., & Hsieh, Y. C. (2007). A lot size model for deteriorating inventory with back-order cancellation. *Springer Lecture Notes in Computer Science*, *4570*, 1002–1011.
6. Nahmias, S., & Smith, S. A. (1994). Optimizing inventory levels in a two-echelon retailer system with partial lost sales. *Management Science*, *40*(5), 582–596.
7. Benjaafar, S., ElHafsi, M., & Huang, T. (2010). Optimal control of a production-inventory system with both backorders and lost sales. *Naval Research Logistics*, *57*(3), 252–265.

Sequencing Problems with Uncertain Parameters and the OWA Criterion

Adam Kasperski and Paweł Zieliński

Abstract In this paper a class of sequencing problems with uncertain parameters is discussed. The uncertainty is modeled by using a discrete scenario set. The Ordered Weighted Averaging (OWA) aggregation operator is used to choose an optimal schedule. The OWA operator generalizes traditional criteria in decision making under uncertainty, such as the maximum, average, median or Hurwicz criterion. In this paper a general framework is proposed and some positive and negative results for a sample problem are presented.

1 Preliminaries

In a sequencing problem, we are given a set of jobs $J = \{J_1, \ldots, J_n\}$ which can be partially ordered by some precedence constraints. The notation $i \rightarrow j$ means that job J_j must be processed after job J_i. We will assume that all the jobs are ready for processing at time 0 and preemption of jobs is not allowed. Thus, a schedule is a feasible permutation of the jobs π which represents an order in which the jobs are processed. We will use Π to denote the set of all feasible schedules. In the classical deterministic case the following parameters for each job J_j can be specified: a nonnegative processing time p_j, a nonnegative due date d_j and a nonnegative weight w_j. We will use $C_j(\pi)$ to denote the completion time of job J_j in schedule

A. Kasperski (✉)
Institute of Industrial Engineering and Management, Wrocław University of Technology, Wrocław, Poland
e-mail: adam.kasperski@pwr.wroc.pl

P. Zieliński
Institute of Mathematics and Computer Science, Wrocław University of Technology, Wrocław, Poland
e-mail: pawel.zielinski@pwr.wroc.pl

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_30,

π. Let $f(\pi)$ be a cost of schedule π. In a deterministic sequencing problem $\mathscr{P}$, we seek a schedule $\pi \in \Pi$ which minimizes the cost function $f(\pi)$.

Suppose that the parameters of the problem are not precisely known. Every possible realization of the problem parameters, denoted by S, is called a *scenario*. We will use $p_j(S)$, $d_j(S)$ and $w_j(S)$ to denote the processing time, due date and weight of job J_j under scenario S, respectively. Without loss of generality we can assume that all these parameters are nonnegative integers. Let *scenario set* $\Gamma = \{S_1, \ldots, S_K\}$ contain all possible, explicitly listed scenarios. Now the job completion time and the cost of schedule π depend on scenario $S \in \Gamma$, and we will denote them by $C_j(\pi, S)$ and $f(\pi, S)$, respectively. Let $v_1, \ldots, v_K$ be numbers such that $v_i \in [0, 1]$, $i \in [K]$, and $v_1 + \cdots + v_K = 1$. Given schedule π, let σ be a permutation of $[K]$ such that $f(\pi, S_{\sigma(1)}) \geq f(\pi, S_{\sigma(2)}) \geq \cdots \geq f(\pi, S_{\sigma(K)})$. The *Ordered Weighted Averaging* aggregation operator (OWA for short) is defined as follows [6]:

$$\mathrm{OWA}(\pi) = \sum_{i \in [K]} v_i f(\pi, S_{\sigma(i)}).$$

In this paper we will study the MIN- OWA $\mathscr{P}$ problem in which we wish to find a schedule $\pi \in \Pi$ minimizing $\mathrm{OWA}(\pi)$. The choice of particular numbers v_i, $i \in [K]$, leads to well known criteria in decision making under uncertainty. Namely, if $v_1 = 1$ and $v_i = 0$ for $i \neq 1$, then we obtain the maximum criterion and the problem is denoted as MIN- MAX $\mathscr{P}$. This is a typical problem considered in robust optimization (see, e.g. [4]). If $v_i = 1/K$ for $i \in [K]$, then we get the average criterion and the problem is denoted as MIN- AVERAGE $\mathscr{P}$. If $v_1 = \alpha$, $v_K = 1 - \alpha$ and $v_i = 0$ for $i = 2, \ldots, K-1$, then we get the Hurwicz pessimism - optimism criterion and the problem is denoted as MIN- HURWICZ $\mathscr{P}$. Finally, when $v_{\lfloor K/2 \rfloor + 1} = 1$ and $v_i = 0$ for $i \neq \lfloor K/2 \rfloor + 1$, then we obtain the median and the problem is denoted as MIN- MEDIAN $\mathscr{P}$.

2 The Maximum Weighted Tardiness Cost Function

Let $T_j(\pi, S) = [C_j(\pi, S) - d_j(S)]^+$ be the *tardiness* of job j in π under scenario S, where $[x]^+ = \max\{0, x\}$. The cost of schedule π under S is the *maximum weighted tardiness*, i.e. $f(\pi, S) = \max_{j \in J} w_j T_j(\pi, S)$. The deterministic counterpart $\mathscr{P}$ is denoted by $1|prec|\max w_j T_j$. We will also discuss a special cases of the problem, with no precedence constraints between jobs and unit job weights, i.e. $1||T_{\max}$.

Theorem 1 MIN- AVERAGE $1||T_{\max}$ *is strongly NP-hard and not approximable within* $7/6 - \varepsilon$ *for any* $\varepsilon > 0$ *unless* $P = NP$; MIN- MEDIAN $1||T_{\max}$ *is strongly NP-hard and not at all approximable unless* $P = NP$.

Proof We show a polynomial time approximation preserving reduction from the MIN k- SAT problem, which is defined as follows. We are given boolean variables

$x_1, \ldots, x_n$ and a collection of clauses $C_1, \ldots, C_m$, where each clause is a disjunction of at most k literals (variables or their negations). We ask if there is a an assignment to the variables that satisfies at most $L < m$ clauses. This problem is strongly NP-hard even for $k = 2$ and its optimization (minimization) version is hard to approximate within $7/6 - \varepsilon$ for any $\varepsilon > 0$ when $k = 3$ (see [2]). Given an instance of MIN 3- SAT, we construct the corresponding instance of MIN- AVERAGE in the following way. We create two jobs J_{x_i} and $J_{\overline{x}_i}$ for each variable x_i. The processing times and weights of all the jobs under all scenarios are equal to 1. The due dates of J_{x_i} and $J_{\overline{x}_i}$ depend on scenario and will take the value of $2i - 1$ or $2i$. We form m scenarios as follows. Scenario S_k corresponds to clause $C_k = (l_1 \vee l_2 \vee l_3)$. For each $q = 1, 2, 3$, if $l_q = x_i$, then the due date of J_{x_i} is $2i - 1$ and the due date of $J_{\overline{x}_i}$ is $2i$; if $l_q = \overline{x}_i$, then the due date of J_{x_i} is $2i$ and the due date of $J_{\overline{x}_i}$ is $2i - 1$; if neither x_i nor $\overline{x}_i$ appears in C_k, then the due dates of J_{x_i} and $J_{\overline{x}_i}$ are set to $2i$. Finally, we fix $v_i = 1/m$ for all $i \in [K]$. Define a subset of the schedules $\Pi' \subseteq \Pi$ such that each schedule $\pi \in \Pi'$ is of the form $\pi = (J_1, J_1', J_2, J_2', \ldots, J_n, J_n')$, where $J_i, J_i' \in \{J_{x_i}, J_{\overline{x}_i}\}$ for $i \in [n]$. Observe that Π' contains exactly 2^n schedules and each such a schedule defines an assignment to the variables such that $x_i = 0$ if J_{x_i} is processed before $J_{\overline{x}_i}$ and $x_i = 1$ otherwise. Assume that the answer to MIN 3- SAT is yes. So, there is an assignment to the variables that satisfies at most L clauses. Consider schedule $\pi \in \Pi'$ which corresponds do this assignment. It is easy to check that if clause C_k is not satisfied, then all jobs in π under S_k are on-time and the maximum tardiness in π under S_k is 0. On the other hand, if clause C_k is satisfied, then the maximum tardiness of π under S_k is 1. In consequence $\frac{1}{K}\sum_{i\in[K]} f(\pi, S_i) \leq L/m$. Assume now that there is a schedule π such that $\frac{1}{K}\sum_{i\in[K]} f(\pi, S_i) \leq L/m$. Notice that $L/m < 1$ by the nonrestrictive assumption that $L < m$. We first show that π must belong to Π'. Suppose that $\pi \notin \Pi'$ and let J_i (J_i') be the last job in π which is not placed properly, i.e. $J_i, (J_i') \notin \{J_{x_i}, J_{\overline{x}_i}\}$. Then J_i (J_i') is at least one unit late under all scenarios and $\frac{1}{K}\sum_{i\in[K]} f(\pi, S_i) \geq 1$, a contradiction. Since $\pi \in \Pi'$ and all processing times are equal to 1 it follows that $f(\pi, S_i) \in \{0, 1\}$ for all $i \in [K]$. Consequently, the maximum tardiness in π is equal to 1 under at most L scenarios and the assignment corresponding to π satisfies at most L clauses.The reduction is approximation-preserving and the inapproximability result immediately holds.

In order to prove the hardness of MIN- MEDIAN $1||T_{\max}$, it is enough to modify the above reduction. Assume first that $L < \lfloor m/2 \rfloor$. We then add to Γ additional $m - 2L$ scenarios with the due dates equal to 0 for all the jobs. So the number of scenarios is $2m - 2L$. We fix $v_{m-L+1} = 1$ and $v_j = 0$ for the remaining scenarios. Now, the answer to MIN 3- SAT is yes, if and only if there is a schedule π whose maximum tardiness is positive under at most $L + m - 2L = m - L$ scenarios. According to the definition of the weights, we have $\mathrm{OWA}(\pi) = 0$. Assume that $L > \lfloor m/2 \rfloor$. We then we add to Γ additional $2L - m$ scenarios with the due dates equal to n for all the jobs. The number of scenarios is then $2L$. We fix $v_{L+1} = 1$ and $v_j = 0$ for all the remaining scenarios. Now, the answer to MIN 3- SAT is yes, if and only if there is a schedule π whose cost is positive under at most L scenarios. The definition of the

weights implies $\mathrm{OWA}(\pi) = 0$. We thus can see that it is NP-hard to check whether there is a schedule π such that $\mathrm{OWA}(\pi) \leq 0$ and the theorem follows. □

We first discuss the minmax version of the problem and extend the polynomial algorithm for the weighted case proposed in [1]. We will use some ideas from [3, 5]. Let us define $F(\pi) = \max_{S\in\Gamma} f(\pi, S)$. We can express the value of $F(\pi)$ as $\max_{j\in J} \max_{S\in\Gamma}[w_j(S)(C_j(\pi, S) - d_j(S))]^+$. Fix a nonempty subset of jobs $D \subseteq J$ and set $F_j(D) = \max_{S\in\Gamma}[w_j(S)(\sum_{i\in D} p_i(S) - d_j(S))]^+$. All job processing times are nonnegative, which implies

$$F_j(D_1) \geq F_j(D_2) \text{ if } D_2 \subseteq D_1 \tag{1}$$

Let $pred(\pi, j)$ be the set of jobs containing job j and all the jobs that precede j in π. Since $C_j(\pi, S) = \sum_{i\in pred(\pi,j)} p_i(S)$, we can rewrite $F(\pi)$ as $F(\pi) = \max_{j\in J} F_j(pred(\pi, j))$. Consider the algorithm shown in the form of Algorithm 1.

Algorithm 1: The algorithm for solving MIN- MAX $1|prec|\max w_jT_j$

```
1 D ← {1, ..., n}, p(S) ← Σ_{j∈D} p_j(S); S ∈ Γ
2 for r ← n downto 1 do
3 |   Find j ∈ D, which has no successor in D and has the minimum value of F_j(D)
4 |   π(r) ← j, D ← D \ {j}; p(S) ← p(S) − p_j(S), S ∈ Γ
5 return π
```

Theorem 2 *Algorithm 1 solves* MIN- MAX $1|prec|\max w_jT_j$ *in* $O(Kn^2)$ *time.*

Proof Let π be the schedule returned by the algorithm. It is clear that π is feasible. Let us renumber the jobs so that $\pi = (1, 2, \ldots, n)$. Let σ be an optimal minmax schedule. Assume that $\sigma(j) = j$ for $j = k + 1, \ldots, n$, where k is the smallest position among all the optimal minmax schedules. If $k = 0$, then we are done, because $\sigma = \pi$ is optimal. Assume that $k > 0$, and so $k \neq \sigma(k) = i$. Let us move the job k just after i in σ and denote the resulting schedule as σ' (see Fig. 1). Schedule σ' is feasible, because π is feasible. We need only consider three cases: (1) If $j \in P\cup R$, then $pred(\sigma', j) = pred(\sigma, j)$ and $F_j(pred(\sigma', j)) = F_j(pred(\sigma, j))$. (2) If $j \in Q \cup \{i\}$, then $pred(\sigma', j) \subseteq pred(\sigma, j)$ and, according to (1), $F_j(pred(\sigma', j)) \leq F_j(pred(\sigma, j))$. (3) If $j = k$, then $F_j(D) \leq F_i(D)$ from the construction of Algorithm 1. Since $pred(\sigma, i) = pred(\sigma', j) = D$, we have $F_j(pred(\sigma', j)) \leq F_i(pred(\sigma, i))$.

From the above three cases we conclude that $F(\sigma') = \max_{j\in J} F_j(pred(\sigma', j)) \leq \max_{j\in J} F_j(pred(\sigma, j)) = F(\sigma)$, so σ' is also optimal, which contradicts the minimality of k. Computing $F_j(D)$ in line 3 of Algorithm 1 requires $O(K)$ time if the sums $p(S)$, $S \in \Gamma$, are used. Thus, the overall running time is $O(Kn^2)$. □

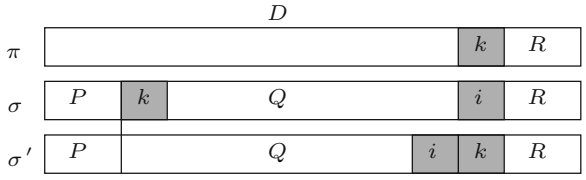


Fig. 1 Illustration of the proof of Theorem 1

Let $f_{\max}$ be an upper bound on the cost of any schedule under any scenario. For example, $f_{\max} = w_{\max}[np_{\max} - d_{\min}]^+$, where $p_{\max}$ is the largest processing time, $w_{\max}$ is the largest weight and $d_{\min}$ is the minimum due date in an input instance.

Theorem 3 MIN- OWA $1|prec|\max w_j T_j$ *is solvable in* $O(f_{\max}^K K n^2)$ *time, which is pseudopolynomial if* K *is constant.*

Proof Let $\boldsymbol{t} = (t_1, \ldots, t_K)$ be a nonnegative integer vector and define $\text{owa}(\boldsymbol{t}) = \sum_{i\in[K]} v_i t_{\sigma(i)}$, where σ is a sequence of $[K]$ such that $t_{\sigma(1)} \geq \cdots \geq t_{\sigma(K)}$. Let $\Pi(\boldsymbol{t})$ be a subset of Π such that $\pi \in \Pi(\boldsymbol{t})$ if $f(\pi, S_i) \leq t_i$ for all $i \in [K]$. Consider the following auxiliary problem: given a vector $\boldsymbol{t}$, check if $\Pi(\boldsymbol{t})$ is not empty and if so, return any schedule $\pi_{\boldsymbol{t}} \in \Pi(\boldsymbol{t})$. This problem can be solved in $O(Kn^2)$ time. Indeed, given $\boldsymbol{t}$, we first form scenario set Γ' by specifying the following parameters for each $S_i \in \Gamma$ and $j \in J$: $p_j(S_i') = p_j(S_i)$, $d_j(S_i') = \max_{C\geq 0} w_j(S_i)(C - d_j(S_i)) \leq t_i$ $(C = t_i/w_j(S_i) + d_j(S_i))$, $w_j(S_i') = 1$. This can be done in $O(Kn)$ time. We then solve the min-max version of the problem for the scenario set Γ' by Algorithm 1 in $O(Kn^2)$ time obtaining schedule π. If $F(\pi)$ over Γ' is 0, then $\pi(\boldsymbol{t}) = \pi$; otherwise $\Pi(\boldsymbol{t})$ is empty. We now show that there exists a vector $\boldsymbol{t}^* = (t_1^*, \ldots, t_K^*)$, where $t_i^* \in \{0, \ldots, f_{\max}\}$, $i \in [K]$, such that each $\pi_{\boldsymbol{t}^*} \in \Pi(\boldsymbol{t}^*)$ minimizes $\text{OWA}(\pi)$. Let π^* be an optimal schedule and let $\boldsymbol{t} = (t_1, \ldots, t_K)$ be a vector such that $t_i = f(\pi^*, S_i)$ for $i \in [K]$. Clearly, $t_i \in \{0, \ldots, f_{\max}\}$ for each $i \in [K]$ and $\pi \in \Pi(\boldsymbol{t})$. By the definition of $\boldsymbol{t}$, we have $\text{owa}(\boldsymbol{t}) = \text{OWA}(\pi^*)$. For any $\pi \in \Pi(\boldsymbol{t})$ it holds $f(\pi, S_i) \leq t_i = f(\pi^*, S_i)$, $i \in [K]$. From the monotonicity of OWA we conclude that each $\pi \in \Pi(\boldsymbol{t})$ must be optimal. The algorithm enumerates all possible vectors $\boldsymbol{t}$ (at most $f_{\max}^K$ vectors) and computes $\pi_{\boldsymbol{t}} \in \Pi(\boldsymbol{t})$ if $\Pi(\boldsymbol{t})$ is nonempty. A schedule $\pi_{\boldsymbol{t}}$ with the minimum value of $\text{owa}(\boldsymbol{t})$ is returned. Hence the problem is solvable in $O(f_{\max}^K K n^2)$. □

Theorem 4 MIN- HURWICZ $1|prec|\max w_j T_j$ *is solvable in* $O(K^2 n^4)$ *time.*

Proof The Hurwicz criterion with $\alpha \in [0, 1]$ can be expressed as $\text{OWA}(\pi) = \alpha \max_{i\in[K]} f(\pi, S_i) + (1-\alpha) \min_{i\in[K]} f(\pi, S_i)$. Let $H_k(\pi) = \alpha \max_{i\in[K]} f(\pi, S_i) + (1-\alpha) f(\pi, S_k)$. Hence $\min_{\pi\in\Pi} \text{OWA}(\pi) = \min_{k\in[K]} \min_{\pi\in\Pi} H_k(\pi)$ and the problem reduces to solving K problems consisting in minimizing $H_k(\pi)$ for a fixed $k \in [K]$. Fix $k \in [K]$ and set $\Pi(t) = \{\pi \in \Pi : f(\pi, S_k) \leq t\} \subseteq \Pi$, where $t \geq 0$. Let $\underline{t}$ be the minimum value of t such that $\Pi(t) \neq \emptyset$. Define $\Psi(t) = \min_{\pi\in\Pi(t)} \max_{i\in[K]} f(\pi, S_i)$, $t \geq \underline{t}$. Hence

$$\min_{\pi\in\Pi} H_k(\pi) = \min_{t\in[\underline{t},\bar{t}]} \alpha\Psi(t) + (1-\alpha)t, \tag{2}$$

where $\bar{t} = \min_{\pi\in\Pi} \max_{i\in[K]} f(\pi, S_i)$, which is due to the fact that $\max_{i\in[K]} f(\pi, S_i) \geq f(\pi, S_k)$. Computing the value of $\Psi(t)$ for a given $t \in [\underline{t}, \bar{t}]$ can be done by a slightly modified of Algorithm 1 in the same running time. It is enough to replace line 3 of Algorithm 1 with the following line: 3' *Find* $j \in D(t)$, *which has no successor in D, and has the minimum value of* $F_j(D)$, *where* $D(t) = \{i \in D : w_i(S_k)[p(S_k) - d_i(S_k)]^+ \leq t\}$. The proof of the correctness of the modified algorithm is almost the same as the proof of Theorem 2. Note that Ψ is a nonincreasing step function on $[\underline{t}, \infty)$, i.e. a constant function on subintervals $[\underline{t}_1, \bar{t}_1) \cup [\underline{t}_2, \bar{t}_2) \cup \cdots \cup [\underline{t}_l, \infty)$, $\bar{t}_{v-1} = \underline{t}_v$, $v = 2, \ldots, l$, $\underline{t}_1 = \underline{t}$. Thus, $\alpha\Psi(t) + (1-\alpha)t$, $\alpha \in (0, 1)$, is a piecewise linear function on $[\underline{t}, \infty)$, a linear increasing function on each subinterval $[\underline{t}_v, \bar{t}_v)$, $v \in [l]$, and attains minimum at one of the points $\underline{t}_1, \ldots, \underline{t}_l$. We now show how these points can be determined. Clearly $\underline{t}_1 = \min_{\pi\in\Pi} f(\pi, S_k)$ and let π_1 be an optimal schedule corresponding to $\Psi(\underline{t}_1)$ computed by the modified Algorithm 1. Let us renumber the jobs so that $\pi_1 = (1, 2, \ldots, n)$. Consider the iteration of the modified algorithm in which job j is placed at the jth position. At this iteration $D = \{1, \ldots, j\}$, $\underline{t}_1$, and j satisfies the condition in line 3'. We can now compute the smallest value of t for which job j violates this condition. In order to do this it suffices to try all values $t_i = w_i(S_k)[p(S_k) - d_i(S_k)]^+$ for $i \in [j-1]$ and fix t_j^* as the smallest among them which violates the condition in line 3' (if the condition holds for all t_i, then $t_j^* = \infty$). Repeating this procedure for each job we get a set of values $t_1^*, \ldots, t_n^*$ and $\underline{t}_2$ is the smallest value among them. The value of $\underline{t}_3$ can be found in the same way. We compute a schedule π_2 corresponding to $\Psi(\underline{t}_2)$ and repeat the previous procedure. Let us now estimate the value of l. Observe that schedule π_i can be obtained from π_{i+1} by decreasing the position of at least one job violating the condition in line 3' (as it must be processed earlier). Furthermore, if the position of such a job in π_{i+1} is k, then its position must be less than k in all schedules $\pi_1, \ldots, \pi_i$. Hence, $l = O(n^2)$ and problem (2) can be solved in $O(Kn^4)$ time and in consequence MIN- HURWICZ $1|prec|\max w_j T_j$ is solvable in $O(K^2 n^4)$ time.

□

Theorem 5 *Suppose that* $v_1 > 0$ *and let* $\hat{\pi}$ *be an optimal solution to the* MIN-MAX $1|prec|\max w_j T_j$ *problem. Then* $\mathrm{OWA}(\hat{\pi}) \leq (1/v_1)\mathrm{OWA}(\pi)$ *for each* $\pi \in \Pi$.

Proof Let σ be a sequence of $[K]$ such that $f(\hat{\pi}, S_{\sigma(1)}) \geq \cdots \geq f(\hat{\pi}, S_{\sigma(K)})$ and ρ be a sequence of $[K]$ such that $f(\pi, S_{\rho(1)}) \geq \cdots \geq f(\pi, S_{\rho(K)})$. It holds $\mathrm{OWA}(\hat{\pi}) = \sum_{j=k}^{K} v_j f(\hat{\pi}, S_{\sigma(j)}) \leq f(\hat{\pi}, S_{\sigma(1)})$. From the definition of $\hat{\pi}$ and the assumption that $v_1 > 0$ we get $f(\hat{\pi}, S_{\sigma(1)}) \leq f(\pi, S_{\rho(1)}) \leq \frac{1}{v_1}\sum_{j\in[K]} v_j f(\pi, S_{\rho(j)}) = \frac{1}{v_1}\mathrm{OWA}(\pi)$. Hence $\mathrm{OWA}(\hat{\pi}) \leq (1/v_1)\mathrm{OWA}(\pi)$. □

Theorem 5 allows us for better approximation of some special cases of the problem. Consider the case of nondecreasing weights, i.e. $v_1 \geq v_2 \geq \cdots \geq v_K$. Since in this case it must hold $v_1 \geq 1/K$, we get that MIN- OWA $1|prec|\max w_j T_j$ is

approximable within a factor not less than K. Finally, we immediately get that MIN-AVERAGE $1|prec|\max w_j T_j$ is approximable within K.

References

1. Aloulou, M. A., & Della, Croce F. (2008). Complexity of single machine scheduling problems under scenario-based uncertainty. *Operations Research Letters*, *36*(3), 338–342.
2. Avidor, A., & Zwick, U. (2002). Approximating min k-sat. In *Algorithms and computation* (pp. 465–475). Berlin: Springer.
3. Kasperski, A. (2005). Minimizing maximal regret in the single machine sequencing problem with maximum lateness criterion. *Operations Research Letters*, *33*(4), 431–436.
4. Kouvelis, P., & Yu, G. (1997). *Robust discrete optimization and its applications*, vol. 14. Berlin: Springer.
5. Volgenant, A., & Duin, C. W. (2010). Improved polynomial algorithms for robust bottleneck problems with interval data. *Computers & Operations Research*, *37*(5), 909–915.
6. Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, *18*, 183–190.

Application of Scheduling Theory to the Bus Evacuation Problem

Corinna Kaufmann

Abstract We consider the problem of scheduling a fleet of buses to evacuate people from a fixed set of gathering points to a fixed set of shelters. In case of uncapacitated shelters this can be modeled as a well-known scheduling problem. Since there are no efficient exact solution methods for this problem, we propose a customized branch-and-bound procedure and compare the performance to a commercial IP solver.

1 Introduction

In the past years a growing number of natural and man-made disasters has made it necessary to develop good algorithms to calculate evacuation plans and estimate evacuation time. In this paper we consider the Bus Evacution Problem (BEP) which was introduced in [3] and considered in a simplified version by [8, 9]. In this paper we consider the special case of the uncapacitated bus evacuation problem (UBEP). To solve this problem we have to schedule a fleet of B buses $\mathscr{B} = \{1, \ldots, B\}$ to evacuate a number of evacuees given in J bus loads $\mathscr{J} = \{1, \ldots, J\}$ out of an incident-affected area. Each bus load j is located at a fixed gathering point s_j of the set $\mathscr{S} = \{1, \ldots, S\}$ and transported along the shortest paths $d_{s_j t}$ to a shelter location t in set $\mathscr{T} = \{1, \ldots, T\}$. Because we consider shortest paths we know that the triangle inequality for the distances holds. To simplify the model, we assume that capacities on shelters are unlimited. This can be motivated by the case of a no-notice evacuation where it is important to get the people out of the affected area as quickly as possible and subsequent transportation between shelters is not as time-critical and therefore not the first objective. There is one bus depot D given and initial traveling times d_{Ds} from the depot to the gathering points $\mathscr{S}$. We will focus on the case with $d_{Ds} = 0$ for all s, but will also indicate the changes necessary for $d_{Ds} > 0$. Buses

C. Kaufmann (✉)
University of Kaiserslautern, Paul-Ehrlich-Straße 14, 67663 Kaiserslautern, Germany
e-mail: kaufmann@mathematik.uni-kl.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_31,

travel alternating between gathering points and shelters and do not travel to several gathering points before going to a shelter. We look for a schedule for each bus such that all people are evacuated and the latest arrival time at a shelter is minimized. We will consider urban instances of the problem, i.e., instances with a large number of evacuees and a small number of gathering points.

We will model the UBEP as the well-known scheduling problem $P_M|s_{ij}|C_{\max}$, i.e., the problem of scheduling N jobs $j = 1, \ldots, N$ on M parallel machines $k = 1, \ldots, M$ with sequence-dependent setup times s_{ij} for $i, j \in \{1, \ldots, N\}$ satisfying the relaxed triangle inequality $s_{ij} \leq s_{ik} + p_k + s_{kj}$, while minimizing the maximum completion time $C_{\max}$. This problem has been extensively studied in the literature. There are two IP formulations developed in [10, 11], as well as a large number of heuristics available (e.g., in [7, 10, 11]). For surveys on scheduling literature concerning setup times see, e.g., [1, 2, 15]. However, to the best of our knowledge, there is no efficient exact solution procedure. Therefore, we will develop a customized branch-and-bound algorithm for the scheduling problem, which will also solve the UBEP. This algorithm will be compared to a commercial IP solver.

In Sect. 2, the UBEP will be modeled as a classical machine scheduling problem. Section 3 gives details on the branch-and-bound algorithm for the scheduling problem as well as improvements for the special case of the urban UBEP. In Sect. 4 computational results are provided and in Sect. 5 further research topics are indicated.

2 Problem Formulation

BEP was shown to be NP-hard in [8]. UBEP can be shown to be NP-hard even in case of just one bus by reduction from the Hamiltonian Path Problem, i.e., the problem of finding a minimum path in a network that contains each node exactly once.

To solve the UBEP as introduced in Sect. 1 we formulate it as a $P_M|s_{ij}|C_{\max}$ scheduling problem. To do so, we assume that bus loads of evacuees are jobs (i.e., each job is associated with a gathering point) and buses are machines. Therefore, we have $N = J$ and $M = B$. Let N_s for $s \in \mathcal{S}$ be the number of bus loads in gathering point s. All evacuees are transported to safety iff each job is processed on one machine. To model the travel time we introduce processing and setup times. A job should be processed by traveling the shortest possible distance from its gathering point s_j to any shelter, i.e., $p_j = \min_{t=1,\ldots,T} d_{s_j,t}$. Since there are likely only a small number of buses, but a large number of evacuees, we assume that $N > M$. Therefore, buses will have to travel back and forth between gathering points and shelters several times. If the job sequence (i, j) is fixed, the best way to do so is to travel the minimum tour from s_i to s_j through an arbitrary shelter t. We will model this as setup times by taking the minimum tour minus the processing time of job i: $s_{ij} = \min_{t=1,\ldots,T} \left(d_{s_i,t} + d_{s_j,t}\right) - p_i$. To model initial traveling times d_{Ds} we introduce a dummy job 0 with $p_0 = 0$ that is processed first on each machine and an initial setup time $s_{0j} = d_{Ds_j}$ for all $j = 1, \ldots, N$.

Since there are IP formulations and many heuristics in the literature (see Sect. 1), we already have a large toolbox for solving this problem. However, there are no efficient exact algorithms. That is why in the next section we introduce a branch-and-bound algorithm for the scheduling problem $P_M|s_{ij}|C_{\max}$ and several improvements that can be used for the urban UBEP because of the special structure of the corresponding instances.

3 Branch-and-Bound Algorithm

In this section the branch-and-bound algorithm for the general scheduling problem $P_M|s_{ij}|C_{\max}$ will be introduced and a few improvements for the special structure of the bus evacuation instances will be presented.

Branching: As was proven in [13] there is an optimal schedule among the list schedules for parallel machine problems with setup-times optimizing a regular objective function if the following allocation rule is applied: The next job in the list is put on the machine where it is finished first. For the branch-and-bound algorithm that means that branching on the set of jobs suffices. Therefore, in each iteration we will introduce at most J new nodes, one for each unscheduled job.

Lower Bounds: For the problem with initial setup times we will use the following lower bound:

$$LB_1 = \frac{1}{M}\left\{\sum_{j=1}^{n}\left[p_j + \min_{i\in\{0,\ldots,N\}} s_{ij}\right]\right\} .$$

This is in fact a corrected version of the lower bound proposed in [11] which had a mistake in it.

In the case without intial setup times, i.e., $s_{0j} = 0$ for all $j = 1, \ldots, n$, this bound is far from optimal. A better lower bound can be obtained by excluding s_{0j} from the calculation of the minimum in LB_1. But then we include M unnecessary setup times because the last job on each machine is not succeeded by a setup time. That means, by subtracting the M largest among the minimum setup times, we obtain a valid lower bound for the problem without initial setup times. Let MS be the set of indices j such that $\min_{i\in\{1,\ldots,N\}} s_{ji}$ is among the M largest setup times. The improved lower bound can be calculated as

$$LB_2 = \frac{1}{M}\left\{\sum_{j=1}^{N}\left[p_j + \min_{i\in\{1,\ldots,N\}} s_{ji}\right] - \sum_{j\in MS} \min_{i\in\{1,\ldots,N\}} s_{ji}\right\} .$$

This can also be easily calculated for a given partial solution: Let l_k, $k = 1, \ldots, M$ be the load scheduled on machine k in the partial schedule. A lower bound

for this partial solution is obtained by calculating the lower bound LB_2 for the set of unscheduled jobs J' and adding it to the average scheduled load $\sum l_k/M$.

Upper Bounds: We will calculate the upper bound by appending the yet unscheduled jobs to the partial list and applying the allocation rule by [13] as for the branching. Several sorting criteria were tested and appending the unscheduled jobs to the partial list in order of non-decreasing average setup-times performed best for most instances. Note that any heuristic based on list scheduling can be used as an upper bound.

Pruning: The search tree can be further reduced by some pruning rules: It is clear that various lists may lead to the same schedule. If in a branching step the allocation rule will put job i on machine k and job j on machine $l \neq k$ we know that when branching from the node of job i in the next step, job j will again be scheduled on $l \neq k$ (since the triangle inequality holds). This means we get two equivalent schedules from the lists $(\ldots, i, j, \ldots)$ and $(\ldots, j, i, \ldots)$ and we can eliminate one of the corresponding branches. We tested a few simple rules that will find such equivalences in reasonable time with little storage requirements. Preliminary experiments showed the following rule to perform best comparing the number of pruned branches against additional computational time and storage requirements:

In the branching step we create M equivalence lists, one for each machine. When branching on one job j which will be scheduled on machine k we will add job j to equivalence list k and store all equivalence lists obtained in this branching step so far except for the k-th in the node of job j. In the following branching step we will prune all branches belonging to jobs in the equivalence list of the corresponding node.

An additional rule can only be applied in case of no initial setup times: In this case the first M jobs will be scheduled as first job on the M machines. Therefore, we require those jobs to be scheduled in lexicographical ordering. All other orders will lead to equivalent schedules where only machines are interchanged.

Improvements for Bus Evacuation Instances For the special structure of UBEP instances we can improve upon the search tree size: In the UBEP setting for urban instances we suppose that there are few gathering points and shelters, but many people, i.e., bus loads to evacuate. All jobs in the same gathering point have the same processing and setup times. Therefore, we will treat them as one job with magnitude N_s. This means, that each node in the tree is branched into at most S new nodes, one for each gathering point, significantly reducing the breadth of the search tree. Furthermore, we will stop branching if in a node there is only one job with magnitude $N_s > 0$. In this case the upper bound will give the optimal solution of this sub-tree.

4 Computational Experiments

The branch-and-bound algorithm presented in Sect. 3 was implemented in C++ and compiled by gcc-Version 4.6.3. Computational performance was compared to the commercial IP solver Cplex using the IP formulation of [11], which was implemented in Python 2.7.3. using Cplex 12.4.0.1. Tests were run on a Dual Intel

Xeon 3.6 GHz processor with 128 GB RAM. 50 randomly generated instances for the general scheduling problem $P_M|s_{ij}|C_{\max}$ with up to 50 jobs were tested. The branch-and-bound algorithm was able to solve 46.00 % of the problems to optimality within a time-limit of 10 min, while Cplex could only solve 22.00 %. The average gap of 57.97 % for Cplex was improved to 1.50 % for the branch-and-bound algorithm.

For the special structure of the UBEP, 90 instances were randomly generated with up to 10 gathering points and 10 shelters similar to the instances in [9]. The branch-and-bound algorithm again clearly outperformed Cplex, solving 48.89 % of the instances to optimality within a time-limit of 10 min, while Cplex could only solve 14.40 %. The average gap of 68.91 % for Cplex was improved to 3.91 % for the branch-and-bound algorithm. Furthermore, performance was compared to that of the branch-and-bound algorithm developed for BEP in [9]. This algorithm solved 38.89 % of the instances to optimality (all of them were also solved to optimality by the new algorithm) and obtained an average gap of 6.94 %. For the instances both algorithms solved to optimality the average runtime of the BEP algorithm was 25.56 s and that of the new algorithm was 6.47 s. These results clearly show that for the special case of uncapacitated shelters the new branch-and-bound algorithm should be favored.

A last test run was made with the five different scenarios of evacuating the city of Kaiserslautern, Germany introduced in [9]. While the branch-and-bound algorithm could solve four of the instances in a 10 min time-limit, three of which in less than 1 s, Cplex could not solve any of the scenarios within the same limit. For the one scenario which the branch-and-bound algorithm could not solve to optimality, the optimality gap was improved from 75.00 % for Cplex to 4.50 % for the branch-and-bound algorithm. Again performance was compared to that of the algorithm in [9]. It also solved 4 of the 5 scenarios to optimality within the given time-limit, but needed an average computation time of 103.10 s compared to 0.73 s for the new algorithm.

5 Conclusion

We modeled the UBEP as the scheduling problem $P_M|s_{ij}|C_{\max}$ and proposed a branch-and-bound algorithm to solve this scheduling problem. For the special structure of the bus evacuation instances improvements for the branch-and-bound algorithm were proposed. Computational results show that this algorithm clearly outperforms the commercial IP solver Cplex and the branch-and-bound algorithm for the capacitated bus evacuation problem developed by [9].

In a next step the proposed algorithm can be extended to solve the capacitated BEP. Also IP formulations and heuristics for the scheduling problem can be adapted to solve the capacitated problem.

Furthermore, this new approach to modeling evacuation problems as well-known scheduling problems offers lots of modeling possibilities: The problem can be modeled with different objective functions such as the total completion time to model

the total evacuation time. It is also possible to include release dates and due dates to model the point in time when evacuees are ready for evacuation and when sick people should arrive at emergency shelters at the latest. Via resource-constrained scheduling it is possible to include constraints of the type: How many police forces or fire-fighters are needed to achieve some target evacuation time or how fast can a given region be evacuated with the help of a given number of rescue forces.

Acknowledgments Partially supported by the Federal Ministry of Education and Research Germany, grant DSS_Evac_Logistic, FKZ 13N12229.

References

1. Allahverdi, A., Gupta, J. N. D., & Aldowaisan, T. (1999). A review of scheduling research involving setup considerations. *Journal of Management Science*, *27*, 219–239.
2. Allahverdi, A., Ng, C. T., Cheng, T. C. E., & Kovalyov, M. Y. (2008). A survey of scheduling problems with setup times or costs. *European Journal of Operational Research*, *187*, 985–1032.
3. Bish, D. (2011). Planning for a bus-based evacuation. *OR Spectrum*, *33*, 629–654.
4. Dunstall, S., & Wirth, A. (2005). A comparison of branch-and-bound algorithms for a family scheduling problem with identical parallel machines. *European Journal of Operational Research*, *167*, 283–296.
5. Elmaghraby, S. E., & Park, S. H. (1974). Scheduling jobs on a number of identical machines. *AIIE Transactions*, *6*, 1–13.
6. França, P. M., Gendreau, M., Laporte, G., & Mller, F. M. (1996). A tabu search heuristic for the multiprocessor scehduling problem with sequence dependent setup times. *International Journal of Production Economics*, *43*, 79–89.
7. Gendreau, M., Laporte, G., & Guimarães, E. M. (2001). A divide and merge heuristic for the multiprocessor scheduling problem with sequence dependent setup times. *European Journal of Operational Research*, *133*, 183–189.
8. Goerigk, M., & Grün, B. (2013). The robust bus evacuation problem. Technical Report, University of Kaiserslautern, Germany.
9. Goerigk, M., Grün, B., & Heßler, P. (2013). *Branch and bound algorithms for the bus evacuation problem*. doi:10.1016/j.cor.2013.07.006.
10. Guinet, A. (1993). Scheduling sequence-dependent jobs on identical parallel machines to minimize completion time criteria. *International Journal of Production Research*, *31*, 1579–1594.
11. Kurz, M. E., & Askin, R. G. (2001). Heuristic scheduling of parallel machines with sequence-dependent set-up times. *International Journal of Production Research*, *39*, 3747–3769.
12. Sarin, S. C., Ahn, S., & Bishop, A. B. (1988). An improved branching scheme for the branch and bound procedure of scheduling n jobs on m parallel machines to minimize total weighte flowtime. *International Journal of Production Research*, *26*, 1183–1191.
13. Schutten, J. M. J. (1996). List scheduling revisited. *Operations Research Letters*, *18*, 167–170.
14. Wesley Barnes, J., & Brennan, J. J. (1977). An improved algorithm for scheduling jobs on identical machines. *AIIE Transactions*, *9*, 25–31.
15. Yang, W. H. (1999). Survey of scheduling research involving setup times. *International Journal of Systems Science*, *30*, 143–155.

Modelling Delay Propagation in Railway Networks

Fabian Kirchhoff

Abstract In this paper we study the accumulation and propagation of delays in (simplified) railway networks. More precisely, we want to estimate the total expected arrival delay of passengers as a cost criterion to be used in a timetable optimisation. Therefore, we want to determine the delay distributions analytically from given source delay distributions. In order to include accumulation and propagation of delays, the source delay distribution must belong to a family of distributions that is closed under appropriate operations. This is the case if we can represent the distribution functions by so called theta-exponential polynomials. A drawback of this representation is the increasing number of parameters needed to describe the results of the operations. A combination with moment approximations allows to solve this problem with sufficient accuracy. Generally, the calculation of propagated delays requires a topological sorting of arrival and departure events. That excludes cyclic structures in the network. We present a relaxation of the topological sorting that allows to (approximately) calculate long run delays in cycles.

1 Related Work

This paper can be assigned to the area of other approaches for the analytical calculation of delay distribution functions. Bueker [2] studied the delay propagation intensively in his dissertation. So far, he hasn't considered the possibility of iterative calculation of delay distributions in cyclic structures of the network. Berger et al. [1] use deterministic delay distributions. In their model there are no cyclic structures.

F. Kirchhoff (✉)
Institute of Applied Stochastics and Operations Research, Erzstr. 1,
38678 Clausthal-Zellerfeld, Germany
e-mail: kirchhoff@math.tu-clausthal.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_32,

2 Modelling Delay Distributions

We use random variables to model source delays and the delays of arrivals and departures in certain stations. First of all, we have to fit the empirical distribution of source delays with sufficient accuracy. We use an approach by Thuemmler et al. [6]. They introduce an algorithm that adapts Hyper-Erlang distributions to given empirical data. We modified their model slightly to be able to fit distributions with $P(X = 0) > 0$, i.e. the event of no delay.

It is a matter of common knowledge [5] that you need three operations to analyse delay propagation in (railway) networks. Let X, Y be two continuous and stochastic independent random variables with $P(X \geq 0) = P(Y \geq 0) = 1$.

1. $Z := X + Y$: $F_Z = F_X * F_Y$ (convolution)
2. $Z := \max\{X, Y\}$: $F_Z = F_X \cdot F_Y$ (multiplication)
3. $Z := \max\{X - s, 0\},\ s \in \mathbb{R}_+$: $F_Z(t) = F_X(t + s) \cdot \mathbf{1}_{[0,\infty)}(t)\ \forall\, t \in \mathbb{R}_+$ (excess beyond)

The first operation is needed if you want to add source delays to current departure or arrival delays. The second operation is used to calculate the delay propagation in connection stations. If you want to subtract buffer times you will use the third operation. It requires the assumption of non-negative delays.

So you need a family of distributions that is closed under these operations. Therefor, we represent delay distribution functions (ddf) as theta-exponential polynomials, introduced in [7]. It is a well-known problem that the complexity of the representation increases very fast. Hence, we need a method to reduce the complexity if it reaches an unacceptable dimension. This is the case if there are numerical problems or too long calculating time.

To reduce the complexity we make use of results by [3, 4]. They searched for solutions for the moment matching problem using Hyper-Erlang distributions with two branches of common order. They showed how to determine the parameters of the Hyper-Erlang distribution. For our purpose this closed-form solution proved to approximate ddfs with sufficient accuracy (Fig. 1).

3 Joined Cycles

In this section we use a graph that is based on a given route network. It represents the relations between feeder and connection lines.

Definition 1 A *route network* is a directed, connected Graph $G = (\mathscr{S}, \mathscr{T})$ with

- $\mathscr{S}$ being the set of vertices representing the stations,
- $\mathscr{T}$ being the set of edges representing the tracks.

Paths in G are called *lines*. The set of all lines is denoted by $\mathscr{L}$. We assume that for all $S \in \mathscr{S}$ there exists a line $L \in \mathscr{L}$ that contains S.

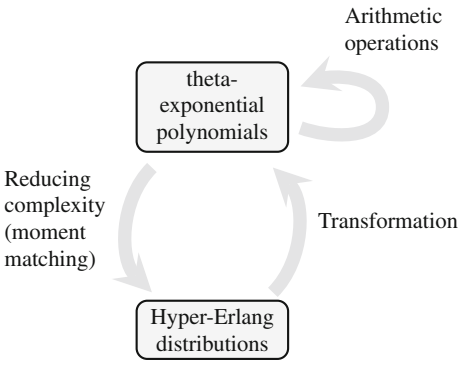


Fig. 1 Analytical calculation of delay distribution functions with the help of complexity reduction

We want to calculate ddfs of all (periodic) arrivals and departures analytically. Therefor we need to order them. The crucial point here are the departures in connection stations.

Definition 2 Let $G = (\mathscr{S}, \mathscr{T})$ be a route network with set of lines $\mathscr{L}$. The corresponding *connection graph* is denoted by $C(G) = (\mathscr{V}, \mathscr{E})$. Let $\mathscr{V}_L = \{V_{L,1}, \ldots, V_{L,n}\}$, $L \in \mathscr{L}$, be an ordered set with the following properties:

- $V_{L,1} \in \mathscr{S}$ represents the first station of line $L \in \mathscr{L}$. $V_{L,n} \in \mathscr{S}$ represents the last station of line $L \in \mathscr{L}$.
- For $1 < i < n$ the elements $V_{L,i} \in \mathscr{S}$ represent the connection stations between first and last station of line $L \in \mathscr{L}$.
- The elements are ordered with respect to their position in the line.

Then we define

$$\mathscr{V} = \{V_L = (V_{L,i}, V_{L,i+1}) \mid V_{L,i}, V_{L,i+1} \in \mathscr{V}_L, L \in \mathscr{L}\}$$

and

$$\mathscr{E} = \big\{(V_L, V_{L'}) \in \mathscr{V} \times \mathscr{V} \mid V_L = (V_{L,i}, V_{L,i+1}), V_{L'} = (V_{L',j}, V_{L',j+1}), \\ V_{L,i+1} = V_{L',j}\big\}.$$

So in the connection graph we merge segments of lines that are located between two connecting stations. This graph provides the information how the delays propagate and how they depend on each other. There is a feasible order for the calculation of the above mentioned arrivals and departures if and only if there exists a topological sorting for the vertices of the connection graph.

Let $\mathscr{K}$ be the set of cycles of a connection graph C and $\mathscr{V}_K$ be the set of vertices of the cycle $K \in \mathscr{K}$. We define $R = \{(K_i, K_j) \in \mathscr{K} \times \mathscr{K} \mid \mathscr{V}_{K_i} \cap \mathscr{V}_{K_j} \neq \emptyset\}$. The transitive closure of R is denoted by R^+. So R contains all pairs of cycles that share at least one vertex. Delays in one of these two cycles influence the delays in the other

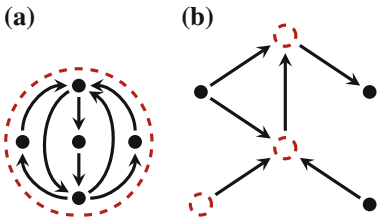


Fig. 2 **a** Joined cycle, **b** joined cycles and singular vertices

cycle. In the following, sets $\mathscr{M} \subset \mathscr{K}$ with $\mathscr{M} = \{K \in \mathscr{K} \mid \exists\, K' \in \mathscr{K} : KR^+ K'\}$ are called *joined cycles*. The set $\mathscr{V}_{\mathscr{M}}$ denotes the disjunct union of all vertices of cycles $K \in \mathscr{M}$. Of course, not all vertices of the connection graph have to be a member of some joined cycle. We call them *singular vertices*. It is easy to verify that there always exists a topological sorting for the set of joined cycles and singular vertices (see Fig. 2). But we still have to order the vertices inside of the joined cycles.

Let us assume that there are cycles in the connection graph. The idea of our approach is the following. In the first step we build the joined cycles (see Fig. 2a). Next we determine the topological sorting of these joined cycles and the singular vertices (see Fig. 2b). The calculation of the ddfs will follow this order. If the next element of the topological sorting is a singular vertex, we will just calculate the ddfs of this vertex. Otherwise, if the next element of the sorting is a joined cycle, we will calculate the ddfs inside of this joined cycle iteratively. In the rest of this paper we assume that we always reach convergence of the ddfs. In fact, this is an open problem.

For the vertices inside of the joined cycles we introduce a pseudo-topological sorting (see Algorithm 2). While (strict) topological sorting allows to visit a vertex only after all of its predecessors have been visited, we visit all vertices in an order respecting the relative number of unvisited predecessors. This number is denoted

Algorithm 1: Pseudo-topological sorting

Input:
- Set of unsorted vertices $\mathscr{V}_u$
- $\forall\, V \in \mathscr{V}_{\mathscr{M}}$: calculated values of α_V
- Empty list of already sorted vertices $\mathscr{L}$

```
1 V_u = V_M;
2 while V_u ≠ ∅ do
3     V_min = {V ∈ V_u | ∀ V* ∈ V_u, V ≠ V* : α_V ≤ α_V*};
4     Choose V ∈ V_min;
5     L ← V;
6     V_u = V_u \ {V};
7     forall the V ∈ V_u do
8         Update α_V;
```

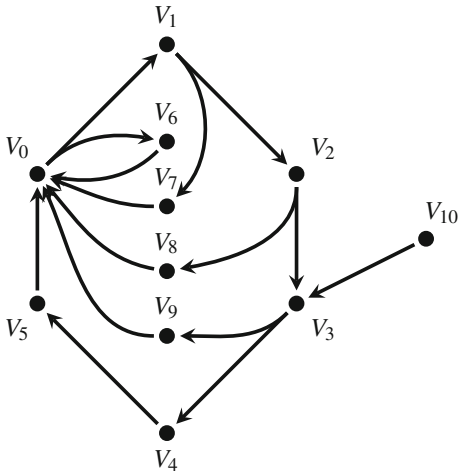


Fig. 3 Connection graph containing 5 cycles, 1 joined cycle and 1 singular vertex

by $\alpha_V \in [0, 1]$, $V \in \mathscr{V}_{\mathscr{M}}$. We start at a vertex with minimal $\alpha_V < 1$. Algorithm 2 requires a vertex V with $\alpha_V < 1$ at the beginning. The reason for this restriction is just the simplification of the pseudocode. Generally, if there isn't a vertex V with $\alpha_V < 1$, i.e. this joined cycle has no predecessors (joined cycle or singular vertex), we start at a vertex with minimal (absolute) number of predecessors. If there are vertices with equal α_V (or equal absolute number of predecessors), the order is chosen at random. Then we mark the chosen vertex as visited. Before we visit the next vertex with minimal α_V, we have to update the values α_V for all unvisited successors of the current vertex.

Generally, in the first step of the iteration we must calculate ddfs of vertices with predecessors whose ddfs haven't been calculated yet. In this case we neglect those predecessors. In the following step of the iteration the ddfs of all vertices have been calculated at least once. Additionally, we test if there is a vertex whose ddfs already reached convergence. In this case we delist it and won't calculate it again in any of the following steps of the iteration.

4 Results

For showing first results we consider the connection graph in Fig. 3. There are 10 vertices. The graph contains 5 cycles. Only vertex V_{10} is not member of a cycle. So V_{10} is the only singular vertex. The vertices $V_1, \ldots, V_9$ are all member of the same joined cycle $\mathscr{M}$. In the first step we determine $(V_{10}, \mathscr{M})$ as the topological sorting of the set of joined cycles and singular vertices. Hence, the ddfs of V_{10} will always be calculated previous to the vertices of $\mathscr{M}$. So we obtain $\alpha_{V_3} = \frac{1}{2}$ and can start algorithm 2.

Table 1 Influence of different sortings on convergence

	Sorting	Average number of calculations (per vertex)
S_1	$V_3, V_4, V_5, V_9, V_8, V_0, V_1, V_2, V_7, V_6$	39.6
S_2	$V_5, V_9, V_8, V_7, V_6, V_0, V_1, V_2, V_3, V_4$	39.6
S_3	$V_0, V_1, V_2, V_3, V_4, V_5, V_9, V_8, V_7, V_6$	40.1
S_4	$V_1, V_3, V_5, V_8, V_6, V_2, V_4, V_9, V_7, V_0$	55.8
S_5	$V_6, V_7, V_8, V_9, V_5, V_4, V_3, V_2, V_1, V_0$	84.7

To test the influence of different sortings we choose a setting (e.g. timetable) that provides

- all vertices the same source delays and buffer times,
- all connections the same buffer times.

The first sorting in Table 1, i.e. S_1, is the pseudo-topological sorting we get by algorithm 2. The idea of sorting S_2 is to calculate the predecessors of the apparently most important vertex V_0 first. Following sorting S_3 means calculating the most important vertex first. S_4 distinguishes form the others in the fact that it doesn't follow the idea of an approximated topological sorting. Instead of that it "jumps" in a way. Sorting S_5 could be considered as the contrary of a topological sorting.

All sortings of Table 1 resulted in convergence of the ddfs. We obtained even the same limiting distributions. But there are differences concerning the speed of convergence. So far, the pseudo-topological sorting does not use any information about the timetable. So it won't be the "best" sorting for all settings. However, in the majority of cases it should be an efficient procedure. Against the background of this work, i.e. optimising timetables with reference to a given network, it could be useful that we don't have to redetermine the sorting for every timetable.

References

1. Berger, A., Gebhardt, A., Mueller-Hannemann, M., & Ostrowski, M. (2011). Stochastic delay prediction in large train networks. *ATMOS*, *20*, 100–111.
2. Bueker, T. (2010). Ausgewaehlte Aspekte der Verspaetungsfortpflanzung in Netzen. Dissertation, RWTH Aachen University.
3. Johnson, M. (1993). Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and erlang distributions. *ORSA Journal on Computing*, *5*, 69–83.
4. Johnson, M., & Taaffe, M. (1989). Matching moments to phase distributions: Mixtures of erlang distributions of common order. *Stochastic Models*, *5*, 711–743.
5. Meester, L. E., & Muns, S. (2007). Stochastic delay propagation in railway networks and phase-type distributions. *Transportation Research*, *41*, 218–230.
6. Thuemmler, A., Buchholz, P., & Telek, M. (2006). A novel approach for phase-type fitting with the EM algorithm. *IEEE Transactions on Dependable and Secure Computing*, *3*, 245–258.
7. Trogemann, G., & Gent, M. (1997). Performance analysis of parallel programs based on directed acyclic graphs. *Acta Informatica*, *34*, 411–428.

Sensitivity Analysis of BCC Efficiency in DEA with Application to European Health Services

Andreas Kleine, Andreas Dellnitz and Wilhelm Rödder

Abstract The CCR model by Charnes et al. [4] on the one hand and BCC model by Banker et al. [3] on the other hand are the most common used approaches of data envelopment analysis (DEA). If we measure efficiency of decision making units (DMUs) by the BCC model, technology is characterized by variable returns to scale. If the inputs and outputs of a DMU are scaled by two parameters such that the BCC (in)efficiency score is unchanged we call this adaptation a bicentric scaling (BS). We introduce a linear program to calculate the BS stability region of all DMUs, efficient or inefficient. Moreover we determine the scale efficiency within the stability region. The new approach is illustrated by a numerical example of European health services. We demonstrate the BS stability region for various states and illustrate consequences on scale efficiency. It is shown that some states can improve scale efficiency without losing BCC efficiency.

1 Introduction

Data envelopment analysis (DEA) measures the relative efficiency of decision making units (DMUs). Charnes et al. [4] introduce a linear program to evaluate the efficiency of DMUs. The so called CCR model assumes constant returns to scales whereas the BCC model by Banker et al. [3] assume variable returns to scale. In the meantime we find numerous extensions and manifold applications of these initial approaches [8, 10].

A. Kleine (✉) · A. Dellnitz · W. Rödder
Fern Universität Hagen, 58084 Hagen, Germany
e-mail: andreas.kleine@fernuni-hagen.de

A. Dellnitz
e-mail: andreas.dellnitz@fernuni-hagen.de

W. Rödder
e-mail: wilhelm.roedder@fernuni-hagen.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_33,

Articles in DEA on sensitivity analysis examine variations of inputs and/or outputs. A stability region is determined within which the efficiency of a specific efficient DMU remains unchanged [5–7, 15]. This paper investigates the sensitivity of BCC efficiency for a given DMU. For a simultaneous shift of inputs and outputs—a bicentric scaling (BS)—we determine a BS stability region such that BCC efficiency is unchanged. Within the BS stability region a DMU can improve or worsen CCR efficiency without losing BCC efficiency.

The article proceeds as follows: The next section briefly summarizes notations of BCC and scale efficiency. Section 3 introduces a bicentric scaling and presents a linear program determining the stability region. Finally Sect. 4 illustrates the approach by a numerical example of the European health service.

2 BCC and Scale Efficiency

A DMU j is characterized by a vector of positive inputs $\mathbf{x}_j = (x_{j1}, \ldots, x_{jm})'$ and outputs $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})'$. All feasible inputs and outputs constitute the production possibility set [12]. The well-known BCC model [3] assumes a technology with variable returns to scale. Applying nonnegative scalars λ_{kj} convex combinations of DMUs build the data envelopment. We focus on input oriented models and thus efficiency of DMU k is calculated by the envelopment form (1) or its dual problem (2). The multiplier form (2) uses input weights $\mathbf{v}_j = (v_{j1}, \ldots, v_{jm})$ and output weights $\mathbf{u}_j = (u_{j1}, \ldots, u_{jn})$. The dual variable u_k corresponds to the convexity restriction:

$$
\begin{array}{l}
\textit{envelopment form} \\
\min \quad \theta_k \\
\text{s.t.} \quad \sum_j \lambda_{kj}\, \mathbf{x}_j \le \theta_k\, \mathbf{x}_k \\
\qquad \sum_j \lambda_{kj}\, \mathbf{y}_j \ge \mathbf{y}_k \\
\qquad \sum_j \lambda_{kj} = 1 \\
\qquad \lambda_{kj} \ge 0\ \forall j,\ \theta_k \text{ free}
\end{array}
\tag{1}
$$

$$
\begin{array}{l}
\textit{multiplier form} \\
\max \quad g_k = \mathbf{u}_k\, \mathbf{y}_k + u_k \\
\text{s.t.} \quad \mathbf{u}_k\, \mathbf{y}_j + u_k - \mathbf{v}_k\, \mathbf{x}_j \le 0 \quad \forall j \\
\qquad \mathbf{v}_k\, \mathbf{x}_k = 1 \\
\qquad \mathbf{u}_k, \mathbf{v}_k \ge 0,\ u_k \text{ free}
\end{array}
\tag{2}
$$

A DMU k is BCC efficient if the optimal solution $\theta_k^* = g_k^* = 1$, i.e. if we do not find a production possibility which dominates activity of DMU k. Increasing (decreasing) returns to scale prevail if all optimal values of u_k^* are positive (negative). If we neglect convexity constraint in envelopment form (1) or set $u_k = 0$ in multiplier form (2) we calculate CCR efficiency [4]. The optimal efficiency score of CCR model is denoted θ_k^{**}.

Since the CCR model assumes constant returns to scale the resulting CCR efficiency is always less or equal than BCC efficiency [12]: $\theta_k^{**} \leq \theta_k^* \leq 1$. A full efficient DMU $\theta_k^{**} = \theta_k^* = 1$ operates in the most productive scale size (mpss) [2].

Scale efficiency is defined by the ratio of CCR and BCC efficiency scores [3]: $SE_k = \theta_k^{**}/\theta_k^* \leq 1$. If efficiencies both coincide a DMU is scale efficient ($SE_k = 1$). In this particular case inefficiencies are due to deviations from the most productive scale size.

3 Sensitivity Analysis with Bicentric Scaling

In the following we vary inputs and outputs of a DMU by radial shifts, a bicentric scaling. Using sensitivity analysis we calculate a BS stability region such that BCC efficiency score is unchanged. In doing so, we are able to identify consequences of bicentric scaling for a given BCC efficiency score. By this means DMUs receive information about potential improvements of their scale efficiency or get indications of undesired change for the worse.

Given an optimal solution of (2) we have

$$g_k^* = \frac{\mathbf{u}_k^* \, \mathbf{y}_k + u_k^*}{\mathbf{v}_k^* \, \mathbf{x}_k} \quad \Longleftrightarrow \quad \mathbf{u}_k^* \, \mathbf{y}_k + u_k^* - g_k^* \, \mathbf{v}_k^* \, \mathbf{x}_k = 0. \tag{3}$$

Next, inputs are adjusted by a shift parameter δ_k and outputs by shift parameter ε_k, respectively. Then DMU k's bicentric scaling forms a trajectory (4) on a hyperplane [11]:

$$\mathbf{u}_k^* \, (1 + \varepsilon_k) \, \mathbf{y}_k + u_k^* - g_k^* \, \mathbf{v}_k^* \, (1 + \delta_k) \, \mathbf{x}_k = 0. \tag{4}$$

Solving and rearranging Eq. (4) yields

$$\varepsilon_k = \frac{\mathbf{u}_k^* \, \mathbf{y}_k + u_k^*}{\mathbf{u}_k^* \, \mathbf{y}_k} \, \delta_k. \tag{5}$$

For a given BCC efficiency the shift parameter ε_k directly depends on δ_k. Hence, we calculate the BS stability region for shift parameter δ_k, only. Applying envelopment form (1) and Eq. (5) we get a linear program (6) for DMU k

$$\begin{array}{ll} \min \,/\, \max & \delta_k \\ \text{s.t.} & \sum_j \lambda_{kj} \, \mathbf{x}_j \leq \theta_k^* \, \mathbf{x}_k \, (1 + \delta_k) \\ & \sum_j \lambda_{kj} \, \mathbf{y}_j \geq \mathbf{y}_k \left(1 + \dfrac{\mathbf{u}_k^* \, \mathbf{y}_k + u_k^*}{\mathbf{u}_k^* \, \mathbf{y}_k} \, \delta_k\right) \\ & \sum_j \lambda_{kj} = 1 \\ & \lambda_{kj} \geq 0 \; \forall j, \; \delta_k \text{ free} \end{array} \tag{6}$$

with optimal values $\delta_k^- = \min \delta_k$ and $\delta_k^+ = \max \delta_k$, respectively.

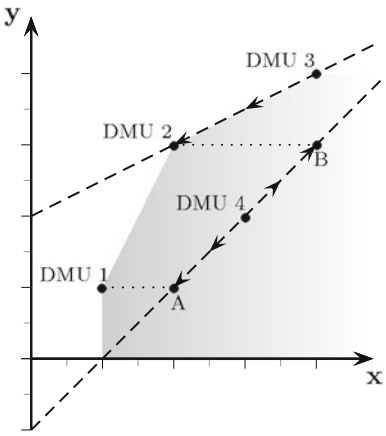


Fig. 1 Bicentric scaling of DMU 3 and DMU 4

The shaded area in Fig. 1 illustrates the production possibility set constituted by DMU 1 to DMU 4 with one input and one output each. The dashed lines represent BCC efficiency hyperplane (4) of DMU 3 and DMU 4. Here, DMU 4 operates under increasing returns to scale (irs) and DMU 3 under decreasing returns to scale (drs). A bicentric scaling of DMU 4 yields point A—with shift δ_4^-—or point B—with shift δ_4^+. In the latter case DMU 4 becomes scale efficient by a bicentric expansion of inputs and outputs. DMU 3 which is BCC efficient yields mpss with a bicentric reduction δ_3^-, i.e. bicentric scaled inputs and outputs match those of DMU 2.

4 Efficiency of European Health Services

In this section bicentric scaling is applied to European health services. We consider health services of 32 European nations—to enhance comparability, only members of European Union or Organization for Economic Cooperation and Development (OECD). Inputs are the number of physicians (phy), nurses (nurs) and beds per 10,000 people. The output is measured by infant survival rate (surv) and live expectancy (exp) (cf. for example [1, 13]). Table 1 summarizes numbers of the countries from World Health Statics 2012 which "represent the best estimates … available in 2011" [14, p. 49].

The results of BCC and CCR efficiency analysis illustrate that six states operate under constant returns to scale (crs) ($\theta_k^* = \theta_k^{**}$, $k = 4, 11, 24, 27, 29, 31$). BS stability region of these mpss countries is fix: $\delta_k^- = \delta_k^+ = 0$ and $SE(\delta_k^-) = SE_k^- = 1 = SE_k^+ = SE(\delta_k^+)$. Some countries are BCC efficient but scale inefficient ($1 = \theta_k^* > \theta_k^{**}$, $k = 13, 15, 16, 19, 28, 32$). For example, Luxembourg ($k = 19$) operates under decreasing returns to scale (drs). A bicentric scaling with an 9.2 % decrease of inputs ($\delta_{19}^- = -0.092$) and a corresponding adaption of outputs increases scale efficiency.

Table 1 BS stability region of European health services

k		Phy	Nurs	Beds	Surv	Exp	θ_k^*	θ_k^{**}	δ_k^-	δ_k^+	SE_k^-	SE_k^+	rts
1	Austria	48.5	78.8	77	249	80	0.515	0.488	−0.065	0.057	1.000	0.908	drs
2	Belgium	30.1	142.0	65	249	80	0.800	0.661	−0.077	0.072	0.884	0.781	drs
3	Bulgaria	37.3	47.0	66	90	74	0.622	0.613	−0.008	0.001	0.905	1.000	irs
4	Cyprus	25.8	43.0	38	332	81	1.000	1.000	0.000	0.000	1.000	1.000	crs
5	Czechia	36.7	87.4	71	332	77	0.626	0.618	−0.228	0.013	0.704	1.000	irs
6	Denmark	34.2	160.9	35	332	79	0.824	0.813	−0.020	0.002	1.000	0.986	drs
7	Estonia	33.3	65.5	54	249	75	0.643	0.627	−0.223	0.020	0.611	1.000	irs
8	Finland	29.1	239.6	62	499	80	0.907	0.869	−0.048	0.000	0.995	0.958	drs
9	France	34.5	80.0	69	332	81	0.748	0.682	−0.114	0.000	1.000	0.911	drs
10	Germany	36.0	111.0	82	332	80	0.681	0.602	−0.146	0.051	1.000	0.851	drs
11	Greece	61.7	36.0	48	332	80	1.000	1.000	0.000	0.000	1.000	1.000	crs
12	Hungary	30.3	64.0	71	199	74	0.630	0.615	−0.142	0.016	0.713	1.000	irs
13	Iceland	37.3	158.8	58	499	82	1.000	0.748	−0.178	0.000	0.899	0.748	drs
14	Ireland	31.7	156.7	49	332	80	0.774	0.734	−0.001	0.051	0.949	0.913	drs
15	Israel	36.5	51.8	35	249	82	1.000	0.903	−0.029	0.000	0.926	0.903	drs
16	Italy	34.9	65.2	36	332	82	1.000	0.924	−0.054	0.000	0.970	0.924	drs
17	Latvia	29.9	48.4	64	124	72	0.640	0.613	−0.050	0.013	0.789	1.000	irs
18	Lithuania	36.1	71.7	68	199	73	0.540	0.521	−0.188	0.023	0.600	1.000	irs
19	Luxembourg	27.7	96.0	56	499	81	1.000	0.919	−0.092	0.000	0.989	0.919	drs
20	Malta	31.1	69.1	45	199	80	0.788	0.665	−0.096	0.059	0.917	0.807	drs
21	Netherlands	28.6	146.0	47	249	81	0.902	0.720	−0.144	0.000	0.909	0.799	drs
22	Norway	41.6	319.3	33	332	81	0.878	0.836	−0.009	0.000	0.959	0.952	drs
23	Poland	21.6	58.0	67	199	76	0.839	0.838	−0.150	0.000	0.484	1.000	irs
24	Portugal	38.7	53.3	33	332	79	1.000	1.000	0.000	0.000	1.000	1.000	crs
25	Romania	22.7	58.8	66	90	73	0.687	0.670	−0.012	0.004	0.900	1.000	irs
26	Slovakia	30.0	66.0	65	142	75	0.560	0.557	−0.083	0.001	0.626	0.998	irs
27	Slovenia	25.1	83.9	46	499	79	1.000	1.000	0.000	0.000	1.000	1.000	crs
28	Spain	39.6	51.1	32	249	82	1.000	0.970	−0.085	0.000	1.000	0.970	drs
29	Sweden	37.7	118.6	28	499	81	1.000	1.000	−0.107	0.000	1.000	1.000	crs
30	Swiss	40.7	164.6	52	249	82	0.857	0.577	−0.164	0.000	0.800	0.673	drs
31	Turkey	15.4	29.0	25	82	75	1.000	1.000	0.000	0.000	1.000	1.000	crs
32	U Kingdom	27.4	101.3	33	199	80	1.000	0.816	0.000	0.000	0.816	0.816	drs

A lot of countries are BCC and CCR inefficient, e.g. Germany ($\theta_{10}^{**} < \theta_{10}^* < 1$) that has the opportunity of achieving scale efficiency. With the optimal scores of (2) for DMU 10

$$\mathbf{v}_{10}^* = (0.0277, 0, 0),\ \mathbf{u}_{10}^* = (0.0003, 0.0353),\ u_{10}^* = -2.2491$$

the BS linear program (6) yields $\delta_{10}^- = -0.1456$ and corresponding $\varepsilon_{10} = -0.034$. Thus, Germany becomes scale efficient ($SE_k^- = 1$) with 14.45 % reduction of inputs and a moderate 3.4 % decrease of outputs. Then BCC and CCR efficiencies coincide, but keep in mind health service of Germany remains still inefficient.

5 Summary

As the example in Sect. 4 illustrates DMUs gain valuable information from bicentric scaling. In addition to traditional analysis DMUs are informed about a numerical measure of adaption. If a DMU does not operate under constant returns to scale then bicentric scaling helps to estimate necessary adaption.

This paper investigates variations of inputs and outputs without loosing BCC efficiency. In addition it is possible to analyze variations of inputs or outputs without changing CCR efficiency [9], so called monocentric scaling. This monocentric scaling results in similar stability regions. This will be subject of further research.

References

1. Alexander, C. A., Busch, G., & Stringer, K. (2003). Implementing and interpreting a data envelopment analysis model to assess the efficiency of health systems in developing countries. *IMA Journal of Management Mathematics*, *14*, 49–63.
2. Banker, R. D. (1984). Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research*, *17*, 35–44
3. Banker, R. D., Charnes, A., Cooper, W. W. (1984). Some models for estimating technical and scale inefficiences in data envelopment analysis. *Management Science*, *30*, 1078–1091.
4. Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *EJOR*, *2*, 429–444.
5. Charnes, A., Cooper, W. W., Lewin, A. Y., Morey, R. C., Rousseau, J. (1985). Sensitivity and stability analysis in DEA. *Annals OR*, *2*, 139–156.
6. Charnes, A., Rousseau, J., Semple, J. (1996). Sensitivity and stability analysis of efficiency classifications in data envelopment analysis. *Journal of Productivity Analysis*, *7*, 5–18.
7. Cooper, W. W., Li, S., Seiford, L. M., Tone, K., Thrall, R. M., Zhu, J. (2001). Sensitivity and stability analysis in DEA: Some recent developments. *Journal of Production Analysis*, *15*, 217–246.
8. Cooper, W. W., Seiford, L. M., & Tone, K. (2007). *Data envelopment analysis—A comprehensive text with models, applications, references and DEA-Solver software*. New York: Springer.
9. Dellnitz, A., Kleine, A., & Rödder, W. (2012). *Ökonomische Interpretationen der Skalenvariablen u in der DEA*. Diskussionsbeitrag der Fakultät für Wirtschaftswissenschaft: FernUniversität in Hagen. 480.
10. Emrouznejad, A., Parker, B. R., & Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholary literature in DEA. *Journal of Socio-Economic Planning Science*, *42*, 151–157.
11. Jahanshahloo, G. R., Hosseinzadeh, F., Shoja, N., Sanei, M., Tohidi, G. (2005). Sensitivity and stability analysis in DEA. *Applied Mathematics and Computation*, *169*, 897–904.
12. Kleine, A. (2004), A general model framework for DEA. *Omega*, *32*, 17–23.
13. Puig-Junoy, J. (1998). Measuring health production performance in the OECD. *Applied Economics Letters*, *5*, 255–259.
14. World Health Organisation. (2012). *World Heath Statistics 2012*. Geneva: WHO Press.
15. Zhu, J. (1996). Robustness of the efficient DMUs in data envelopment analysis. *EJOR*, *90*, 451–460.

Competition for Resources

The Equilibrium Existence Problem in Congestion Games

Max Klimm

Abstract Congestion games are an elegant model to study the effects of selfish usage of resources. In my thesis—of the same title as this note—we characterized the maximal conditions for which the existence of a pure Nash equilibrium can be guaranteed for four variants of congestion games: weighted congestion games, congestion games with resource-dependent demands, congestion games with variable demands, and bottleneck congestion games. This note reviews the main results obtained there.

1 Introduction

Infrastructure networks are the lifelines of our civilization. From the first irrigation systems in Egypt and Mesapotamia, over the Roman roads and the first railways to the latest generation of fiber-optic network cables—infrastructure systems continue to have a tremendous impact on the fortunes of humankind. A common characteristic of infrastructure networks is that they are used by a large number of selfish individuals who strive to minimize their private cost of using the network rather than optimizing the global state of the system. Such interactions between selfishly acting individuals are modeled and analyzed with the theory of *non-cooperative games*.

Rosenthal [18] proposed a particularly elegant and simple model to study the effects of selfish resource usage which he called *congestion games*. In such a game, we are given a finite set of resources. Each player is associated with a set of feasible allocations, where each allocation is a subset of the resources, and strives to choose an allocation so as to minimize the sum of the cost of all resources used. The cost of a resource depends on the number of players using that resource and is given as

M. Klimm (✉)
Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136,
10623 Berlin, Germany
e-mail: klimm@math.tu-berlin.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_34,

a resource-specific function of the demand for that resource. An allocation vector constitutes a pure Nash equilibrium if no player can decrease her cost by a unilateral deviation. Rosenthal proved that each such a game always possesses a Nash equilibrium in pure (i.e., deterministic) strategies. This is a remarkable result since for general games by John F. Nash's famous theorem c.f. [16] only a Nash equilibrium in mixed (i.e., randomized) strategies is guaranteed.

Congestion games model a variety of strategic interactions, most prominently traffic in street networks. Here, the set of resources corresponds to the set of street segments of a network. Each player is associated with an origin and a destination in the network, and her set of feasible allocations corresponds to the set of paths from the origin to the destination. The cost functions on the resources are used to model transit times which typically increase as the usage of a street segment increases.

2 Weighted Congestion Games

While obviously important, congestion games do not take into account that the users may contribute to a different extent to the congestion on the resources. In traffic networks, e.g., a truck clearly contributes more to the utilization of the street than a regular car. Such interactions are captured more realistically by *weighted congestion games*. In such a game, each player has a strictly positive *demand* that she places on the chosen resources and the cost of each resource is a function of the aggregated demand of all players using that resource. Hence, unweighted congestion games are a special case of weighted congestion games in which all players have unit demand. In contrast to unweighted congestion games, weighted congestion games may fail to admit a pure Nash equilibrium c.f. [7, 8, 15]. On the positive side, it is known that for affine resource cost functions or exponential resource cost functions a pure Nash equilibrium always exists c.f. [7, 11, 17].

These positive results establish the existence of a pure Nash equilibrium *independent* of the underlying structure of the game, i.e., independent of the number of players, the combinatorial structure of their strategies, and their demands. Such independence is desirable because the number of players and their types (expressed in terms of their demands and their strategies) are only known to the players and subject to frequent changes. Thus, it is natural to study the existence of equilibria with respect to the cost functions of the resources.

It was an open problem which maximal sets of cost functions guarantee the existence of a pure Nash equilibrium in weighted congestion games. In my thesis [14, Chap. 3], we give a complete answer to this question. Specifically, we show that a set $\mathscr{C}$ of continuous cost functions guarantees the existence of a pure Nash equilibrium in all weighted congestion games if and only if one of the following two cases holds: (i) $\mathscr{C}$ only contains affine functions; (ii) $\mathscr{C}$ only contains exponential cost functions with the property that there is a constant $\phi \in \mathscr{R}$ and for each $c \in \mathscr{C}$ two constants $a_c, b_c \in \mathscr{R}$ such that $c(x) = a_c e^{\phi x} + b_c$ for all $x \geq 0$. The necessity of these conditions is even valid for games with three players. This implies in particular that

for every non-affine and non-exponential function c there is a three-player weighted congestion game where all resources have cost function c and that does not possess a pure Nash equilibrium.

We provide a similar characterization for two-player weighted congestion games. Here, a set $\mathscr{C}$ of continuous cost functions guarantees the existence of a pure Nash equilibrium in all two-player weighted congestion games if and only if $\mathscr{C}$ contains only monotonic functions and each two non-constant functions $c_1, c_2 \in \mathscr{C}$ are linear transformations of each other, i.e., there are $a, b \in \mathscr{R}$ such that $c_1(x) = a\, c_2(x) + b$ for all $x \geq 0$.

These characterizations precisely explain under which maximal conditions a pure Nash equilibrium is guaranteed to exists. Thus, they may help to predict and explain unstable traffic distributions in infrastructure networks. In telecommunication networks, e.g., relevant cost functions are the so-called $M/M/1$-delay functions and in road networks frequently used functions are monomials of degree four put forward by the US Bureau of Public Roads. Our characterizations imply, that for these types of cost functions, there is *always* an instance with three players and identical cost functions that is unstable in the sense that a pure Nash equilibrium does not exist. On the other hand, our characterizations can be used to design a stable system: e.g., uniform $M/M/1$-delay functions are consistent for two-player games. For the formal statements and the proofs of these results for weighted congestion games, see also [10].

3 Congestion Games with Resource-Dependent Demands

In a weighted congestion game, each player has a *unique* demand that she places on all the resources contained in her strategy. Dropping the assumption that the demand of a player is equal for all resources we obtain *congestion games with resource-dependent demands*. Among others, such games may be used to yield a much more accurate model of traffic networks that incorporates the fitness of different vehicle types w.r.t. the physical properties of road segments, such as slopes, terrain, and so on. Although congestion games with resource-dependent demands allow to model a much broader scope of applications than weighted congestion games, they have not received a similar attention in the literature, in the past. Most previous work concentrated on the special case of *scheduling games* where each allocation consists of exactly one resource e.g. [2, 3, 5].

There are two natural ways of defining the players' private cost functions. In the first variant, called *proportional games*, the cost of the resources is interpreted as a monetary per-unit cost. In this regime, it is natural to assume that each player incurs a cost equal to the sum of the cost of the used resources multiplied with her respective demand. We also study a slightly different class of games, called *uniform* games. They differ from proportional games solely in the fact that in the definition of the players' private cost, the cost of the resources is *not* multiplied with the player's

demands. Such cost structure occurs when the resource cost is interpreted as latencies or travel times and thus are the same for each user, regardless of their demands.

Building on the results obtained for weighted congestion games, in my thesis [14, Chap. 4], we give a complete characterization of the existence of pure Nash equilibria in these games. Specifically, a set $\mathscr{C}$ of continuous cost functions guarantees the existence of pure Nash equilbria in proportional games if and only if $\mathscr{C}$ only contains affine functions. Furthermore, $\mathscr{C}$ guarantees the existence of a pure Nash equilibrium in uniform games if and only if $\mathscr{C}$ only contains constant functions. This characterization is even valid for three-player games. For the formal statements and the proofs of these results for congestion games with resource-dependent demands, see also [9].

4 Congestion Games with Variable Demands

Although congestion games with resource-dependent demands capture the main features of many interesting applications, they do not take into account the elasticity of the demand due to price changes. Such elasticity is an intrinsic property of many applications, most prominently the flow control problem in telecommunication networks. Here, the players strive to establish an unsplittable data stream in a network. The sending rate is reduced if the latency increases and is increased if the latency decreases. To model elasticity of demands, in my thesis [14, Chap. 5], we initiate the study of *congestion games with variable demands*. It is assumed that each player is associated with an interval of feasible demands and a non-decreasing and concave utility function modeling the utility received from satisfying a certain demand. In every strategy profile, each player chooses both a feasible demand and exactly one feasible subset of resources. The private payoff of each player then is defined as the difference between the utility received from the chosen demand and the cost incurred on the used resources.

As before, we obtain a complete characterization of the existence of pure Nash equilibria in congestion games with variable demands in the proportional and uniform cost model, respectively. Specifically, we prove that a set $\mathscr{C}$ of continuous and non-negative cost functions guarantees the existence of a pure Nash equilibrium in proportional congestion games with variable demands if and only if at least one of the following two cases holds: (i) there is a constant $\phi > 0$ and for each $c \in \mathscr{C}$ a constant $a_c > 0$, such that $c(x) = a_c\, e^{\phi x}$ for all $x \geq 0$; (ii) for each $c \in \mathscr{C}$ there are constants $a_c > 0$ and $b_c \geq 0$ such that $c(x) = a_c\, x + b_c$ for all $x \geq 0$. In addition, we prove that $\mathscr{C}$ is consistent for uniform congestion games with variable demands if and only if (i) holds. For the formal statements and the proofs of these results for congestion games with variable demands, see also [9].

5 Bottleneck Congestion Games

So far, it is assumed that the players strive to minimize the *sum* of the cost on the chosen resources. In many scenarios, however, sum-objectives do not represent the players' incentives correctly. An example of such a situation is data streaming in telecommunication networks where the delay of a data stream is restricted by the available bandwidth of the links on the chosen path. The total delay experienced by a selfish user is closely related to the performance of the link with the least bandwidth. To capture this situation more realistically, Banner and Orda [4] introduced *bottleneck congestion games*. They differ from weighted congestion games in the fact that in each strategy profile the private cost of each player is the *maximum* of the cost of all chosen resources. Banner and Orda proved the existence of a pure Nash equilibrium for non-decreasing cost functions on the resources.

In my thesis [14, Chap. 6], we generalize the existence result of Banner and Orda by weakening the assumptions on the cost functions, assuming that the cost of the resources may even depend on the *set* of players using them. Even for these more general cost functions, we prove the existence of a *strong equilibrium* for this class of bottleneck congestion games with set-dependent cost. Strong equilibria are a strengthening of the pure Nash equilibrium concept that is even resilient to deviations of coalitions of players that decrease the private cost of each of its members. Each strong equilibrium is a pure Nash equilibrium, but not conversely.

Further, we study *splittable* bottleneck congestion games. In such a game, each player is associated with a strictly positive demand that she is allowed to split fractionally over the sets of resources available to her. For continuous and non-decreasing cost functions on the resources, we show that splittable bottleneck congestion games admit a strong equilibrium.

The existence of strong equilibria in bottleneck congestion games raises some important questions regarding the computability of equilibria in such games. While for unweighted congestion games with sum-objective the complexity of computing pure Nash equilibria is relatively well understood [1, 6], the complexity status of computing equilibria under bottleneck-objectives remained open. In my thesis [14, Chap. 7], we give first results in this direction.

First, we propose a generic algorithm that computes a strong equilibrium for any unweighted bottleneck congestion game with non-decreasing cost functions. This algorithm crucially relies on a *strategy packing oracle* that decides for a given vector of resource capacities whether there exists a strategy profile that obeys the capacity constraint on each resource, and that outputs such a strategy profile if it exists. The running time of this algorithm is essentially determined by the running time of the oracle. This implies that the problem of computing a strong equilibrium in an unweighted bottleneck congestion game with non-decreasing cost can be reduced to solving the strategy packing problem. As a characterization, we prove the converse direction, i.e., solving a strategy packing problem is reducible to computing a strong equilibrium in an unweighted bottleneck congestion game with non-decreasing cost.

There are a number of important classes of bottleneck congestion games for which a strategy packing oracle can be implemented in polynomial time, including single-commodity networks, branchings, and matroids. In all these cases, a strong equilibrium can be determined efficiently using the generic algorithm. For general games, however, we show that the computation of a strong equilibrium is NP-hard. This holds even for two-commodity networks. For unweighted bottleneck congestion games with single-commodity network or matroids strategies we show an interesting dichotomy. Although for both classes of games an efficient algorithm calculating a strong equilibrium exists, the recognition of a strong equilibrium is coNP-hard. For the formal statements and the proofs of these results for bottleneck congestion games, see also [12, 13].

References

1. Ackermann, H., Röglin, H., & Vöcking, B. (2008). On the impact of combinatorial structure on congestion games. *Journal of the ACM*, *55*(6), 1–22.
2. Andelman, N., Feldman, M., & Mansour, Y. (2009). Strong price of anarchy. *Games and Economic Behavior*, *65*(2), 289–317.
3. Awerbuch, B., Azar, Y., Richter, Y., & Tsur, D. (2006). Tradeoffs in worst-case equilibria. *Theoretical Computer Science*, *361*(2–3), 200–209.
4. Banner, R., & Orda, A. (2007). Bottleneck routing games in communication networks. *IEEE Journal on Selected Areas in Communications*, *25*(6), 1173–1179.
5. Even-Dar, E., Kesselman, A., & Mansour, Y. (2007). Convergence time to Nash equilibrium in load balancing. *ACM Transactions on Algorithms*, *3*(3), 1–21.
6. Fabrikant, A., Papadimitriou, C., Talwar, K. (2004). The complexity of pure Nash equilibria. In *Proceedings of the thirty-sixth annual ACM symposium on theory of computing*, pp. 604–612.
7. Fotakis, D., Kontogiannis, S., & Spirakis, P. (2005). Selfish unsplittable flows. *Theoretical Computer Science*, *348*(2–3), 226–239.
8. Goemans, M., Mirrokni, V., Vetta, A. (2005). Sink equilibria and convergence. In *Proceedings of 46th annual IEEE symposium on foundations of computer science*, pp. 142–154.
9. Harks, T., Klimm, M. (2011). Congestion games with variable demands. In *Proceedings of the 13th conference on theoretical aspects of rationality and knowledge*, pp. 111–120.
10. Harks, T., & Klimm, M. (2012). On the existence of pure Nash equilibria in weighted congestion games. *Mathematics of Operations Research*, *37*(3), 419–436.
11. Harks, T., Klimm, M., & Möhring, R. (2011). Characterizing the existence of potential functions in weighted congestion games. *Theory of Computing Systems*, *49*(1), 46–70.
12. Harks, T., Hoefer, M., Klimm, M., & Skopalik, A. (2012a). *Computing pure Nash and strong equilibria in bottleneck congestion games*. To appear: Mathematical Programming.
13. Harks, T., Klimm, M., & Möhring, R. (2012b). Strong equilibria in games with the lexicographical improvement property. *International Journal of Game Theory*, *42*(2), 461–482.
14. Klimm, M. (2012). Competition for resources: The equilibrium existence problem in congestion games. PhD thesis.
15. Libman, L., & Orda, A. (2001). Atomic resource sharing in noncooperative networks. *Telecommunication Systems*, *17*(4), 385–409.
16. Nash, J. (1950). Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences*, *36*, 48–49.
17. Panagopoulou, P., & Spirakis, P. (2006). Algorithms for pure Nash equilibria in weighted congestion games. *Journal of Experimental Algorithmics*, *11*, 1–19.
18. Rosenthal, R. (1973). A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, *2*(1), 65–67.

Measurement of Risk for Wind Energy Projects: A Critical Analysis of Full Load Hours

André Koukal, Stefan Lange and Michael H. Breitner

Abstract In scientific literature, profitability analyses of on- and offshore wind energy projects and assessments of general conditions for such projects usually make use of the full load hours (FLH) key figure to determine the annually produced energy. They also serve for the calculation of the project value and other financial key figures. This procedure leads to accurate results if only the expected value of each parameter is taken into account. However, it is difficult to choose an adequate type of distribution and to define suitable distribution parameters for the FLH when project risks are considered. In this paper, a different approach using the more basic parameter of the average wind speed and a Weibull distribution in combination with the technical availability and other discounts is provided. It aims at estimating the annual electricity generation by simultaneously taking uncertainties into account. This approach is integrated into a discounted cash flow (DCF) model on which a Monte Carlo simulation is applied. Finally, a case study for a fictitious offshore wind park in the German North Sea is conducted. It is shown that the application of the presented approach leads to more precise distributions of the outcomes than the standard analysis with FLH.

1 Introduction and Research Background

The development of renewable energy technologies and in particular the wind energy as a major part of it has been increasingly furthered by governments in various countries in order to expand electricity generation capacities and also reduce greenhouse

A. Koukal (✉) · S. Lange · M. H. Breitner
Leibniz Universität Hannover, Hannover, Germany
e-mail: koukal@iwi.uni-hannover.de

S. Lange
e-mail: stefan.lange89@gmail.com

M.H. Breitner
e-mail: breitner@iwi.uni-hannover.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_35,

gas emissions. The establishment and improvement of methods to assess the economic potential and to quantify risks of certain projects are required to support these targets. In previous research of Madlener et al. [3] and Koukal and Breitner [4] the FLH key figure is used to determine the generated electricity of an offshore wind park. Respective values for a specific region are taken from publicly available reports. These values are discounted to consider shadowing effects, technical availability, and other effects and are integrated into a DCF model. This approach leads to accurate results when only the expected values of financial key figures are assessed within the model.

In order to consider project risks, they extend their financial models with probability distributions for every risky parameter in combination with an application of a Monte Carlo simulation (MCS). As they discuss, this approach results only in a rough approximation of the distribution of every target key figure. However, the assignment of a specific probability distribution for the FLH key figure is especially critical for multiple reasons:

1. The FLH key figure is a highly aggregated measure that combines various aspects with different levels of influence on the measure.
2. Any of the aggregated aspects has a different probability distribution that describes the specific risk.
3. Due to the inhomogeneous influences and the diverse probability distributions of the individual aspects it is even harder to set up a suitable probability distribution for the aggregated FLH key figure.

2 Estimating the Generated Electricity

To avoid the difficulties of setting up probability distributions for the FLH key figure, we discard this measure and consider all previously aggregated aspects separately. In detail, we use a sequence of the aspects which results in an electricity generation chain of a wind park (Fig. 1).

Wind speed and Weibull distribution In current literature, the Weibull distribution is used to describe the distribution of wind speed which is the most basic parameter of the energy production chain. The Weibull distribution function is presented by

$$P(v < v_i < v + d_v) = P(v > 0) \left(\frac{k}{c}\right) \left(\frac{v_i}{c}\right)^{k-1} * exp\left[-\left(\frac{v_i}{c}\right)^{k}\right] dv, \quad (1)$$

where c is the scale factor (in m/s), k is the unitless Weibull shape factor that varies between 1 and 4, v is the wind speed, v_i is a particular wind speed and d_v is an incremental wind speed [5]. The relationship between the average wind speed and the two Weibull parameters is given by

$$\bar{v} = c * \Gamma\left(1 + \frac{1}{k}\right), \quad (2)$$

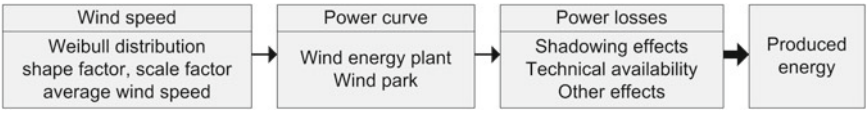


Fig. 1 Electricity generation chain after discarding the FLH key figure

with $\bar{v}$ as the average wind speed and $\Gamma()$ as the gamma function [5]. The scale factor c can be determined by a transformation of the Eq. 2 by

$$c = \frac{\bar{v}}{\Gamma\left(1 + \frac{1}{k}\right)} \tag{3}$$

In order to measure the amount of energy generated by a wind park, the Weibull parameters have to be determined first. One approach uses the wind power density (WPD) [6]. This measure indicates how energetic the winds are. As long time series of wind data exist for many regions this approach suits best to approximate the values of the Weibull parameters. The WPD is defined by

$$WPD = \frac{1}{2}\rho\bar{v}^3, \tag{4}$$

with the WPD measured in W/m^2, $\bar{v}$ as the average wind speed and ρ as the air density, which can be approximated by

$$\rho = 1.225 - (1.1194 \times 10^{-4}) \times z, \tag{5}$$

with z as the location's elevation above sea level in m. To determine the Weibull parameters, the average WPD is derived from an observed data set, e.g. for one year. Next, this value is matched to the average WPD from the Weibull distribution by varying the scale factor. Based on the Weibull distribution, the time at each wind speed within a specified period is described by

$$h_i = \frac{t_i}{T}, \tag{6}$$

with T as the total time of an observed period, e.g. 8,760 h for one year, and t_i as the number of occurrences of wind speed i within the observed period.

Power curve Every type of wind energy plant has its individual power curve that describes the power output for any wind speed. The same applies to an entire wind park. Multiplying the number of occurrences of one wind with the respective power output results in the electricity generation at any wind speed i, which is determined by

$$E_i = h_i * P_i * T, \tag{7}$$

where P_i is the power output of the wind energy plant or wind park. In order to get the theoretical maximum electricity generation of a wind energy plant or wind park for an observed period, the sum of the electricity generation of every wind speed i has to be calculated by

$$E_{total} = \sum E_i = T * \sum h_i * P_i \tag{8}$$

Energy losses After the theoretical maximum electricity production is determined, additional factors have to be considered. Shadowing effects reduce the average power output in dependence of the distances between the individual wind energy plants [7]. Time lags are the second factor that discount the generated electricity. They result from the huge area of a wind park with varying wind speeds within the park [7]. The third discount is the technical availability. Faulstich et al. [1] outline a bathtub curve for the failure rates of wind turbines that results in different availability levels over time. They outline high failure rates in the early life period, lower and constant failure rates in the useful life period and increasing failure rates at the end of the life-cycle in the "wearout period". The technical availability is the complementary probability of the failure rate. The fourth discounts are losses based on the transfer of the electricity. The efficiency of voltage converters and cables affect this discount.

Probability distributions for risk measurement The approach of setting up probability distributions for every risky key figure applied by [3] and [4] is retained but modified due to the replacement of the FLH key figure. To consider uncertainties regarding the wind speed, the average wind speed parameter of the Weibull distribution is defined as risky parameter. A normal distribution can be derived from the historical wind data and is assigned to this parameter. While shadowing effects, time lags, and electricity transfer losses are assumed to be fixed, a normal distribution is applied to the yearly values of the technical availability.

3 Case Study: Offshore Wind Park in the German North Sea

Our case study is based on the research of Koukal and Breitner [4]. We use their DCF model with the same assumptions for every parameter and probability distribution, except the FLH key figure. This key figure is replaced by the aspects of the electricity generation chain presented in Sect. 2.

To derive Weibull parameters we use wind data from the FINO1 research platform [2] from a period of 9 years. We estimate an average wind power density of 955 w/m^2, an average wind speed of 9.85 m/s, a Weibull shape factor of $k = 2,36$ and a scale factor $c = 11,12$ m/s. To consider discounts of shadowing effects and time lags, we follow the argumentation of [7] and apply discounts of 5 and 6 %. The technical availability starts with 80 % [1] and increases in the early life period to 85 % after 5 years. In the wearout period after 10 years of operation it decreases with 1 % in every following year.

Table 1 Comparison of key results

	Mean	Std. Dev.	95 % percentile	Prob. CF ≥ 0	Kurtosis
Model with electricity generation chain	€ 23.9 m	€43.1 m	€-47.5 m	71.2 %	3.00
Model with FLH key figure [4]	€ 96.7 m	€67.0 m	€-34.9 m	87.8 %	2.57

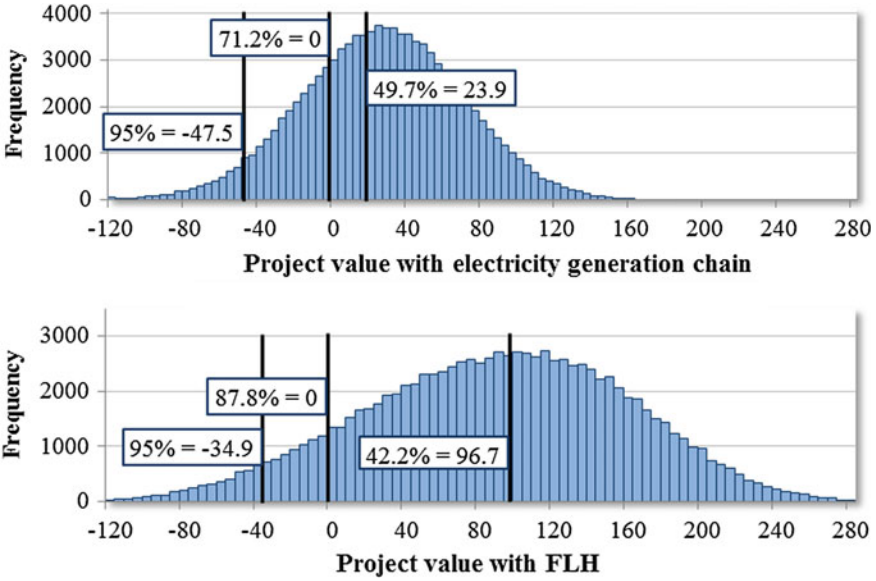


Fig. 2 Distribution of the project value (millions of € , 100,000 simulations)

3.1 Results

All assumptions about the electricity generation chain are implemented in a modified cash flow model of [4]. The key results are presented in Table 1 while the distribution of the project value is presented in Fig. 2.

The results allow several statements about the project and the applied model:

1. Our new approach results in a lower mean value than the old model with FLH. However, both approaches result in positive expected project values.
2. The standard deviation is lower although more input factors are considered and thus, possible values are more concentrated around the mean value.
3. The kurtosis of the project value distribution with the new approach is higher. More results are in the tails of the distribution and more extreme, but rare events are projected

4. The project value at a 95 % confidence level is negative in both cases but significantly lower with the new approach.

4 Discussion, Limitations and Conclusion

While previous research of [3] and [4] draws a pretty favorable conclusion with regards to corporate and project finance of offshore wind parks, our findings illustrate that respective projects are more risky than previously anticipated. The lower mean value of € 23.9m in combination with the project value at a 95 % confidence level of only −47.5 m indicates that the use of FLH results in an overestimation of returns an investor can achieve. The aggregation of several factors with different inherent risks and probability distributions to the FLH key figure leads to distortions of results. There is no linear connection between the average wind speed and the FLH. Thus, using the Weibull distribution function and technical availability in order to derive the probability distribution of the project value generally provides a more realistic approach.

However, there are some limitations. Firstly, the average wind speed and the Weibull shape factor are used as input variables. While it is possible to derive these values from historical wind data for many regions, it might be difficult to get adequate data for any possible location. Nevertheless, it is still possible to use approximate values and estimations. Secondly, there are few approximations about the technical availability and other limiting factors, e.g. the bathtub curve and the applied standard deviation for the technical availability. In general, the usage of more input variables increases the chance of errors. However, further research needs to be conducted in order to verify whether the results of the introduced method with the electricity generation chain are still more realistic compared to simply using FLH.

It can be concluded that the approach presented in this paper helps to improve the consideration and measurement of risks when assessing wind energy projects. The replacement of the aggregated FLH key figure by the wind speed, its distribution, and other factors offers a more detailed analysis and leads to more specific results.

References

1. Faulstich, S., Hahn, B., & Tavner, P. J. (2011). Wind turbine downtime and its importance for offshore deployment. *WIND ENERGY*, *14*, 327–337.
2. Forschungs- und Entwicklungszentrum Fachhochschule Kiel GmbH: FINO1 - Forschungsplattformen in Nord- und Ostsee Nr. 1 (2013) http://www.fino1.de/en/location-sea-floor-waves-wind Cited 20 Mai 2013.
3. Madlener, R., Siegers, L., & Bendig, S. (2009). Risikomanagement und -controlling bei Offshore-Windenergieanlagen. *Zeitschrift für Energiewirtschaft*, *33*, 135–146.

4. Koukal, A., Breitner, M.H.: Decision Support Tool for Offshore Wind Parks in the Context of Project Financing. In: Proceedings of the International Annual Conference of the German Operations Research Society (OR2012), Hannover (2012).
5. Seguro, J. V., & Lambert, T. W. (2000). Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis. *Journal of Wind Engineering and Industrial Aerodynamics*, *85*(1), 75–84.
6. Troen, I., Lundtang Petersen, E.: European Wind Atlas. Roskilde, Denmark (1989).
7. Jarass, L., Obermair, G. M., Voigt, W.: Windenergie: zuverlssige Integration in die Energieversorgung. Springer (2009).

An Integer Programming Approach to the Hospitals/Residents Problem with Ties

Augustine Kwanashie and David F. Manlove

Abstract The classical Hospitals/Residents problem (HR) models the assignment of junior doctors to hospitals based on their preferences over one another. In an instance of this problem, a stable matching M is sought which ensures that no blocking pair can exist in which a resident r and hospital h can improve relative to M by becoming assigned to each other. Such a situation is undesirable as it could naturally lead to r and h forming a private arrangement outside of the matching. The original HR model assumes that preference lists are strictly ordered. However in practice, this may be an unreasonable assumption: an agent may find two or more agents equally acceptable, giving rise to *ties* in its preference list. We thus obtain the Hospitals/Residents problem with Ties (HRT). In such an instance, stable matchings may have different sizes and MAX HRT, the problem of finding a maximum cardinality stable matching, is NP-hard. In this paper we describe an Integer Programming (IP) model for MAX HRT. We also provide some details on the implementation of the model. Finally we present results obtained from an empirical evaluation of the IP model based on real-world and randomly generated problem instances.

1 Introduction

The Hospital Residents Problem (HR) has applications in a number of centralised matching schemes which seek to match graduating medical students (residents) to hospital positions. Examples of such schemes include the National Resident Matching Program (NRMP) in the US [1], and the Scottish Foundation Allocation Scheme

Supported by Engineering and Physical Sciences Research Council grant EP/K010042/1.

A. Kwanashie (✉) · D. F. Manlove
School of Computing Science,University of Glasgow, Glasgow, UK
e-mail: a.kwanashie.1@research.gla.ac.uk

D. F. Manlove
e-mail: David.Manlove@glasgow.ac.uk

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_36,

(SFAS), which ran until recently in Scotland.. The challenges presented by these and other applications have motivated research in the area of algorithms for matching problems.

Formally an instance I of HR involves a set $R = \{r_1, r_2, ..., r_{n_1}\}$ of *residents* and $H = \{h_1, h_2, ..., h_{n_2}\}$ of *hospitals*. Each resident $r_i \in R$ ranks a subset of H in strict order of preference with each hospital $h_j \in H$ ranking a subset of R, consisting of those residents who ranked h_j, in strict order of preference. Each hospital h_j also has a capacity $c_j \in \mathbb{Z}^+$ indicating the maximum number of residents that can be assigned to it. A pair (r_i, h_j) is called an *acceptable pair* if h_j appears in r_i's preference list and r_i on h_j's preference list. A *matching* M is a set of acceptable pairs such that each resident is assigned to at most one hospital and the number of residents assigned to each hospital does not exceed its capacity. A resident r_i is *unmatched* in M if no acceptable pair in M contains r_i. We denote the hospital assigned to resident r_i in M as $M(r_i)$ (if r_i is unmatched in M then $M(r_i)$ is undefined) and the set of residents assigned to hospital h_j in M as $M(h_j)$. A hospital h_j is *under-subscribed* in M if $|M(h_j)| < c_j$. An acceptable pair (r_i, h_j) can *block* a matching M or forms a *blocking pair* with respect to M if r_i is either unmatched or prefers h_j to $M(r_i)$ and h_j is either under-subscribed or prefers r_i to at least one resident in $M(h_j)$. A matching M is said to be *stable* if there exists no blocking pair with respect to M.

We consider a generalisation of HR which occurs when the preference lists of the residents and hospitals are allowed to contain *ties*, thus forming the Hospital/Residents Problem with Ties (HRT). In an HRT instance a resident (hospital respectively) is indifferent between all hospitals (residents respectively) in the same tie on its preference list. In this context various definitions of stability exist. We consider *weak stability* [2] in which a pair (r_i, h_j) can *block* a matching M if r_i is either unmatched or strictly prefers h_j to $M(r_i)$ and h_j is either under-subscribed or strictly prefers r_i to at least one resident in $M(h_j)$. A matching M is said to be *weakly stable* if there exists no blocking pairs with respect to M. Henceforth we will refer to a weakly stable matching as simply a stable matching.

Every instance of the HRT problem admits at least one stable matching. This can be obtained by breaking the ties in both sets of preference lists in an arbitrary manner, thus giving rise to a HR instance which can then be solved using the Gale-Shapley algorithm for HR [3]. The resulting stable matching is then stable in the original HR instance. However, in general, the order in which the ties are broken yields stable matchings of varying sizes [4] and the problem of finding a maximum weakly stable matching given an HRT instance (MAX HRT) is known to be NP-hard [4]. Various approximation algorithms for MAX HRT can be found in the literature [5, 6] with the best current algorithm achieving a performance guarantee of 3/2.

Due to the NP-hardness of MAX HRT and the need to maximize the cardinality of stable matchings in practical applications, *Integer Programming* (IP) can be used to solve MAX HRT instances to optimality. This paper presents a new IP model for MAX HRT (Sect. 2). In Sect. 3 we provide some details on the implementation of the model. Finally Sect. 4 summarises some of the results obtained by evaluating the model against real-world and randomly generated problem instances. Proofs and more detailed empirical results can be found in [7].

2 An IP Model for MAX HRT

In this section we describe an IP model for MAX HRT which is a non-trivial extension of an earlier IP model for a 1–1 restriction of MAX HRT due to Podhradskỳ [8]. Let I be an instance of HRT consisting of a set $R = \{r_1, r_2, ..., r_{n_1}\}$ of residents and $H = \{h_1, h_2, ..., h_{n_2}\}$ of hospitals. We denote the binary variable $x_{i,j}$ $(1 \leq i \leq n_1, 1 \leq j \leq n_2)$ to represent an acceptable pair in I formed by resident r_i and hospital h_j. Variable $x_{i,j}$ will indicate whether r_i is matched to h_j in a solution or not: if $x_{i,j} = 1$ in a given solution J then r_i is matched to h_j in M (the matching obtained from J), else r_i is not matched to h_j in M. We define $rank(r_i, h_j)$, the rank of h_j on r_i's preference list, to be $k + 1$ where k is the number of hospitals that r_i strictly prefers to h_j. An analogous definition for $rank(h_j, r_i)$ holds. Obviously for HRT instances agents in the same tie have the same rank. We define $rank(r_i, h_j) = rank(h_j, r_i) = \infty$ for an unacceptable pair (r_i, h_j). With respect to a pair (r_i, h_j), we define the set $T_{i,j} = \{r_p \in R : rank(h_j, r_p) \leq rank(h_j, r_i)\}$ and $S_{i,j} = \{h_q \in H : rank(r_i, h_q) \leq rank(r_i, h_j)\}$. We also define the set $P(r_i)$ to be the set of hospitals that r_i finds acceptable and $P(h_j)$ to be the set of residents that h_j finds acceptable. The resulting model is presented below. Constraint 1 ensures that each resident is matched to at most one hospital and Constraint 2 ensures that each hospital does not exceed its capacity. Finally Constraint 3 ensures that the matching is stable by ruling out the existence of any blocking pair.

$$\max \sum_{i=1}^{n_1} \sum_{h_j \in P(r_i)} x_{i,j}$$

subject to

1. $$\sum_{h_j \in P(r_i)} x_{i,j} \leq 1 \quad (1 \leq i \leq n_1)$$
2. $$\sum_{r_i \in P(h_j)} x_{i,j} \leq c_j \quad (1 \leq j \leq n_2)$$
3. $$c_j \left(1 - \sum_{h_q \in S_{i,j}} x_{i,q}\right) - \sum_{r_p \in T_{i,j}} x_{p,j} \leq 0 \quad (1 \leq i \leq n_1, h_j \in P(r_i)) x_{i,j} \in \{0, 1\}$$

Theorem 1 *Given a HRT instance I modeled as an IP using model1, a feasible solution to model1 produces a weakly stable matching in I. Conversely a weakly stable matching in I corresponds to a feasible solution to model1.*

3 Implementing the Model

In this section we describe some techniques used to reduce the size of the HRT model generated and improve the performance of the IP solver. Techniques were described in [9] for removing acceptable pairs that cannot be part of any stable matching from HRT instances with ties on one side of the preference lists only. The *hospitals-offer* and *residents-apply* algorithms described identify pairs that cannot be involved in any stable matching, nor form a blocking pair with respect to any stable matching, and remove them from the instance. This produces a reduced HRT instance that would yield fewer variables and constraints when modelled as an IP thus speeding up the optimisation process. The original instance and the reduced instance have the same set of stable matchings.

A number of steps were taken to improve the optimisation performance of the models. These include placing a lower bound on the objective function and providing an initial solution to the CPLEX solver. Both can be obtained by executing any of the approximation algorithms [9] on the HRT instance (the 3/2-approximation algorithm for HRT with ties on one side only due to Király [10] was chosen).

4 Empirical Evaluations

An empirical evaluation of the IP model was carried out. Large numbers of random instances of HRT were generated by varying certain parameters relating to the construction of the instance and passed on to the CPLEX IP solver. Data from past SFAS matching runs were also modelled and solved. This section discusses the methodology used and some of the results obtained. Experiments were carried out on a Linux machine with 8 Intel(R) Xeon(R) CPUs at 2.5 GHz and 32 GB RAM.

Although the theoretical model has been proven to be correct, it is still important to verify the correctness of the implementation. The system was tested to ensure a high degree of confidence in the results obtained. The correctness of the pre-processing steps and the IP solution were evaluated by generating multiple instances (100,000) of various sizes (with up to 400 residents) and testing the stability and size of the resulting matching against both the original and the trimmed problem instance. For all the instances tested, the solver produced optimal stable matchings.

Random HRT problem instances were generated. The instances consist of n_1 residents, n_2 hospitals and C posts where n_1, n_2 and C can be varied. The hospital posts were randomly distributed amongst the hospitals. Other properties of the generated instance that can be varied include the lengths of residents' preference lists as well as a measure of the density t_d of ties present in the preference lists. The tie density t_d ($0 \leq t_d \leq 1$) of the preference lists is the probability that some agent is tied to the agent next to it in a given preference list. At $t_d = 1$ each preference lists would be contained a single tie while at $t_d = 0$ no tie would exist in the preference lists of all agents thus reducing the problem to an HR instance. We define the size of the instance as the number of residents n_1 present.

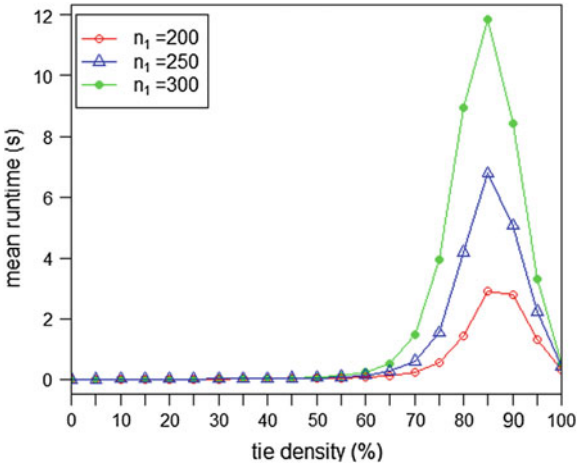


Fig. 1 Mean runtime versus t_d

Since ties cause the size of stable matchings to vary, an obvious question to investigate is how the variation in tie density affects the runtime of the IP model and the size of the maximum stable matchings found. These values were measured for multiple instances of MAX HRT while varying the tie density t_d of hospitals' preference lists. This was done for increasing sizes ($n_1 = 200, 250, 300$) of the problem instance with the residents' preference list being kept strictly ordered at 5 hospitals each. A total of 10,000 instances were randomly generated for each tie density value (starting at $t_d = 0\,\%$ to $t_d = 100\,\%$ with an interval of $5\,\%$) and instance size. For each instance $C = n_1$ and $n_2 = \lfloor 0.07 \times nR \rfloor$.

To avoid extreme outliers skewing the mean measures, we define what we regard as a reasonable solution time (300 s) and abandon search if the solver exceeds this cut-off time. In most cases this cut-off was not exceeded: in [7] we show the percentage of instances that were solved before the cut-off was exceeded for the values of n_1 and t_d considered (the lowest of which was 97.76 %).

From Figs. 1 and 2 we see that the mean and median runtime remain significantly low for instances with $t_d < 60\,\%$ but then steeply increase until they reach their peaks (in the region of 80–90 %) before falling as the tie density approaches 100 %. From a theoretical perspective, it is known that the problem is polynomially solvable when the tie density is at both 0 and 100 % and it is easy to see how the IP solver will find these cases trivial. As the tie density increases the number of stable matchings that the instance is likely to admit also increases, explaining the observed increase in the runtime. The *hospitals-offer* and *residents-apply* algorithms used to trim the instance also play their part in this trend with limited trimming done for higher tie densities.

We also looked to answer the question of how scalable the IP model is by increasing n_1 and measuring the mean and median time taken to solve multiple random

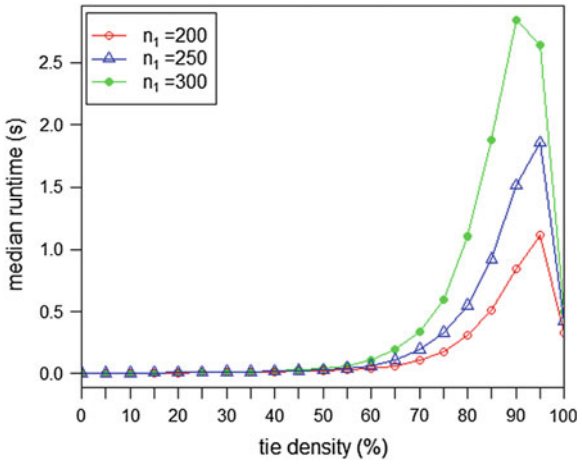


Fig. 2 Median runtime versus t_d

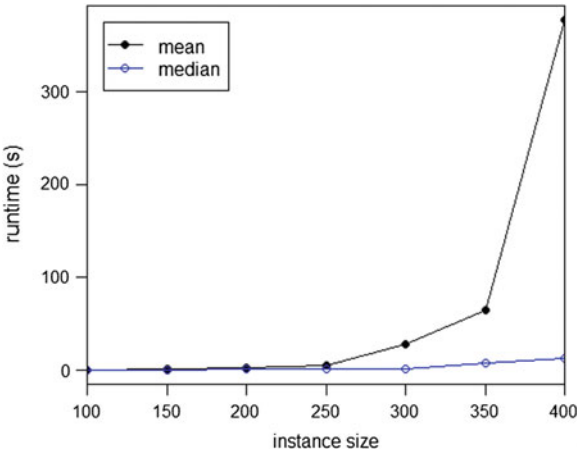


Fig. 3 Mean and median runtime versus n_1

instances. The tie densities of the hospitals' preference lists were set to 0.85 on all instances. The instance size n_1 was increased by 50 starting at $n_1 = 100$. A total of 100 instances for each instance size was generated. The number of hospitals n_2 in each instance was set to $\lfloor 0.07 \times n_1 \rfloor$. No cut-off was set for this experiment. Figure 3 shows how the mean computational time increases with n_1. We assume the increasingly sharp difference between the mean and median is due to the presence of outliers due to exceptionally difficult instances.

Another question worth asking is whether the IP model can handle instance sizes found in real-world applications. In [9], various approximation algorithms and heuristics were implemented and tested on real datasets from the SFAS matching scheme

Table 1 SFAS IP results

Year	n_1	n_2	t_d (%)	Time (s)	$\|M\|$	$\|M'\|$ from [9]
2006	759	53	92	92.96	758	754
2007	781	53	76	21.78	746	744
2008	748	52	81	75.50	709	705

for 2006, 2007 and 2008 where the residents' preferences are strictly ordered with ties existing on the hospitals' preference lists. With the IP model, it is now possible to trim the instances using the techniques mentioned in Sect. 3, generate an optimal solution and compare the results obtained with those reported in [9]. Results from these tests showed that, while some algorithms did marginally better than others, all the algorithms developed generated relatively large stable matchings with respect to the optimal values. Table 1 shows this comparison where M' denotes the largest stable matching found over all the algorithms tested in [9].

References

1. National Resident Matching Program website. http://www.nrmp.org.
2. Irving, R. W. (1994). Stable marriage and indifference. *Discrete Applied Mathematics*, *48*, 261–272.
3. Gale, D., & Shapley, L. S. (1962). College admissions and the stability of marriage. *American Mathematical Monthly*, *69*, 9–15.
4. Manlove, D. F., Irving, R. W., Iwama, K., Miyazaki, S., & Morita, Y. (2002). Hard variants of stable marriage. *Theoretical Computer Science*, *276*(1–2), 261–279.
5. Király, Z. (2012). Linear time local approximation algorithm for maximum stable marriage. In *Proceedings of MATCH-UP*, Vol. 12, pp. 99–110.
6. McDermid, E. (2009). A 3/2 approximation algorithm for general stable marriage. In *Proceedings of ICALP 09*. LNCS (Vol. 5555, pp. 689–700). Berlin: Springer.
7. Kwanashie A., Manlove D. F. (2013) An integer programming approach to the hospitals/residents problem with ties. Technical Report 1308.4064, Computing Research Repository, Cornell University Library.
8. Podhradskỳ, A. (2011). Stable marriage problem algorithms. Masters thesis, Masaryk University, Faculty of Informatics.
9. Irving, R. W., & Manlove, D. F. (2009). Finding large stable matchings. *ACM Journal of Experimental Algorithmics*, *14*, 2.
10. Király Z. (2008) Better and simpler approximation algorithms for the stable marriage problem. *Algorithms-ESA 2008* (pp. 623–634). Berlin: Springer.

Learning in Highly Polarized Conflicts

Sigifredo Laengle and Gino Loyola

Abstract Negotiations are often conducted in highly polarized environments, which are also uncertain and dynamic. However, the intense rivalry involved in these conflicts does not always prevent an agreement from being reached. A recently proposed static model sets out the conditions under which either an agreement is achieved or negotiations break down in this environment [4]. Nevertheless, important aspects related to partial mutual knowledge of players in a dynamic context are not yet been studied. To fill this gap, we develop an extension of the static game to modelling highly polarized conflicts in an uncertain, asymmetric and dynamic environment. In this extension both parties bargain multiple negotiation rounds under uncertain threats that are materialised only if an agreement is not reached. If a negotiation breakdown occurs, each party learns about these threats from the outcome observed in the previous round. This paper presents the most important results, and a short discussion about possible applications. In particular, we provide the conditions that characterise different paths for negotiations held under polarized environments, which matches the observed evolution of many of these conflicts in the real world.

1 Introduction

Since the pioneering work of [5], two important questions remain open in distributive bargaining theory. First, what are the conditions under which a negotiation breakdown can emerge as an outcome. Second, in case of reaching an agreement, what and how source of bargaining power of parties shape the properties of the deal.

S. Laengle (✉) · G. Loyola
Department of Management Control, Universidad de Chile, Diagonal Paraguay 257, Santiago de Chile, Chile
e-mail: slaengle@fen.uchile.cl

G. Loyola
e-mail: gloyola@fen.uchile.cl

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_37,

Indeed, in the original game proposed by Nash (the so-called Nash demand game), whereas disagreement is not possible, there is a multiplicity of equilibria that implies a continuum of agreements. Subsequent literature on negotiation games has explored both issues, but in general through the high sophistication of the original game or the proposition of a new game at all. In fact, this literature has offered answers to these questions by either introducing perturbations into the original game, adopting an incomplete information environment, or using a dynamic approach to model a distributive bargaining situation.

In an attempt to address both issues under an even simpler approach, a model of distributive negotiation was constructed [4] in which bargainers suffer a negative externality proportional to the surplus captured by their rival. The paper does an extensive analysis of the related literature (which are not cited here) and examines the impact of negative externalities on equilibrium properties of the classic Nash demand game.

Nevertheless, important aspects related to partial mutual knowledge of players in a dynamic context are not yet been studied. To fill this gap, we develop an extension of the static game to modelling highly polarized conflicts in an uncertain, asymmetric and dynamic environment. In this extension both parties bargain multiple negotiation rounds under uncertain threats that are materialised only if an agreement is not reached. If a negotiation breakdown occurs, each party learns about these threats from the outcome observed in the previous round.

Related contributions to our work are [2, 3], which also study the role played by externalities among several buyers in negotiations held with a seller in a multi period context, but the base model has several differences to our model. Other interesting contribution is [1], but it differs to our work that the first-mover player is the seller. Under this set-up, it is shown that large enough negative externalities can be a source of bargaining *delays*, irrespective of the existence of a deadline. The general framework of this literature is not able, however, to yield a complete bargaining disagreement, or alternatively, an indefinite delay.

This paper presents the most important mathematical results. In particular, we provide the conditions that characterise different paths for negotiations held under polarized environments, which matches the observed evolution of many of these conflicts in the real world. Section 2 presents a generalized static model of [4], which is extended to the dynamic bargaining under uncertainty in Sect. 3.

2 An Extended Static Model

Let us considerer the following **distributive negotiation game with externalities** extended from [4]. It consists of a 2-agent non-cooperative game with players named as Emil and Frances. Emil and Frances choose simultaneously strategies x, and y respectively. The agents try to maximize the function utility given by $f(x, y) \doteq x - \gamma_E y$ and $g(x, y) \doteq y - \gamma_F x$ for Emil and Frances respectively, where $\gamma_E, \gamma_F \geq 0$

are constants, which represent the **externalities**. Let $\pi \geq 0$ be the **pie** to be distributed among the players.

Let u and v be constants, which represent the **outside opportunities** of Emil and Frances respectively, which satisfy $-\gamma_E \pi \leq u \leq \pi$, $-\gamma_F \pi \leq v \leq \pi$. We say that the strategy pair $(\overline{x}, \overline{y})$ is a **Nash equilibrium** of the game if and only if

$$f(\overline{x}, \overline{y}) = \max_x \{f(x, \overline{y}) : x \geq 0, x + \overline{y} \leq \pi, \text{ and } x - \gamma_E \overline{y} \geq u\};$$

and

$$g(\overline{x}, \overline{y}) = \max_y \{g(\overline{x}, y) : y \geq 0, \overline{x} + y \leq \pi, \text{ and } y - \gamma_F \overline{x} \geq v\}.$$

Theorem 1 *If* $\pi\gamma_E\gamma_F + (u(1+\gamma_F) + v(1+\gamma_E)) \leq \pi$, *then the strategies pair* $(\overline{x}, \overline{y})$ *is a Nash equilibrium satisfies* $\overline{x} + \overline{y} = \pi$,

$$(u + \gamma_E \pi)/(1+\gamma_E) \leq \overline{x} \leq (\pi - v)/(1+\gamma_F), \text{ and}$$
$$(v + \gamma_F \pi)/(1+\gamma_F) \leq \overline{y} \leq (\pi - u)/(1+\gamma_E).$$

Proof If (x, y) is a Nash equilibrium strategy pair,[1] then x is an optimum of Emil's optimization problem, given y, and must therefore satisfy the KKT conditions. Moreover, since the problem is concave, x is a global optimum. In other words, there exist $\lambda_{1E}, \lambda_{2E}, \lambda_{3E} \in \mathbb{R}$ that satisfy the following properties: (1) stationarity:

$$\frac{d}{dx}(f(x, y) + \lambda_{1E}x - \lambda_{2E}(x+y-\pi) + \lambda_{3E}(x - \gamma_E y - u)) = 1 + \lambda_{1E} - \lambda_{2E} + \lambda_{3E} = 0;$$

(2) primal feasibility: (2a) $x \geq 0$, (2b) $x + y \leq \pi$ and (2c) $x - \gamma_E y \geq u$; (3) dual feasibility: $\lambda_{1E}, \lambda_{2E}, \lambda_{3E} \geq 0$; and (4) complementary slackness: (4a) $\lambda_{1E} x = 0$, (4b) $\lambda_{2E}(x + y - \pi) = 0$ and (4c) $\lambda_{3E}(x - \gamma_E y - u) = 0$. Analogously, y is an optimum for Frances' optimization problem, given x. There then exist multipliers that satisfy equations (1) through (4) above in which the corresponding substitutions have previously been made (subscript F for E, y for x and x for y). We now compute the strategies x, y that satisfy the Nash equilibrium. **First**, let us assume that $\lambda_{2E} = 0$. Then $\lambda_{1E} + \lambda_{3E} = -1$ by (1), which contradicts (3). Thus, λ_{2E} must be not equal to 0, therefore $x + y = \pi$ by (4b). **Second**, the equality $x + y = \pi$ and the condition (2c) $x - \gamma_E y \geq u$ prove that x and y must satisfy the conditions $x \geq (\gamma_E \pi + u)/(1+\gamma_E)$ and $y \leq (\pi - u)/(1+\gamma_E)$ respectively. Likewise in the Frances' case, since $\lambda_{2F} \neq 0$ implies $x + y = \pi$, thus $y \geq (\gamma_F \pi + v)/(1+\gamma_F)$ and $x \leq (\pi - v)/(1+\gamma_F)$. In other words, x must be in the closed interval $[(u + \gamma_E \pi)/(1+\gamma_E), (\pi - v)/(1+\gamma_F)]$ and y in the closed interval $[(v + \gamma_F \pi)/(1+\gamma_F), (\pi - u)/(1+\gamma_E)]$. Given these results, it is not difficult to prove that both interval strategy are not empty (by calculating the difference between the upper bound and the lower bound of both intervals) if only if $\pi\gamma_E\gamma_F + (u(1+\gamma_F) + v(1+\gamma_E)) \leq \pi$. **Third**, let us note that the lower bound

[1] We write x for $\overline{x}$ and y for $\overline{y}$ for simplifying the notation.

of x must be between 0 and π, which is equivalent to $-\gamma_E\pi \leq u \leq \pi$. The same equivalence is obtained from the upper bound of y. Likewise, in the case of the lower bound of y an the upper bound of x, which are each one equivalent to condition $-\gamma_F\pi \leq v \leq \pi$. □

3 A Dynamic and Uncertain Bargaining Process

3.1 Bargaining Under Uncertainty

Asymmetric Information Let us suppose that the pie is $\pi \doteq 1$, and the Frances's outside opportunity v is a random variable that takes two possible values $0 < v_H \leq 1$ (high) and $v_L \doteq 0$ (low) with probability θ and $1-\theta$ respectively (with $\theta \in]0, 1[$). The Emil's outside opportunity is $u \doteq 0$. Furthermore, Frances observes her outside opportunity (v) and the Emil's outside opportunity (u) completely. Instead, Emil does not observe the Frances's outside opportunity, but he observes his outside opportunity completely.

Emil chooses a strategy x, and y_H, y_L respectively for Frances. The agents try to maximize the function utility given by $f(x, y_H, y_L) \doteq x - \gamma_E(\theta y_H + (1-\theta)y_L)$ and $g(x, y_H, y_L) \doteq \theta(y_H - \gamma_F x) + (1-\theta)(y_L - \gamma_F x)$ for Emil and Frances respectively.

We say that the strategies tuple $(\overline{x}, \overline{y}_H, \overline{y}_L)$ is a Nash equilibrium of the game if and only if

$$f(\overline{x}, \overline{y}_H, \overline{y}_L) \doteq \max_x \big\{ f(x, \overline{y}_H, \overline{y}_L) : x \geq 0, x + \theta\overline{y}_H + (1-\theta)\overline{y}_L \leq 1, \text{ and } x - \gamma_E(\theta\overline{y}_H + (1-\theta)\overline{y}_L) \geq 0 \big\};$$

and

$$g(\overline{x}, \overline{y}_H, \overline{y}_L) \doteq \max_{y_H, y_L} \big\{ g(\overline{x}, y_H, y_L) : y_H, y_L \geq 0, \overline{x} + y_H \leq 1, \overline{x} + y_L \leq 1, y_H - \gamma_F\overline{x} \geq v_H, \text{ and } y_L - \gamma_F\overline{x} \geq 0 \big\}.$$

To solve this problem, we apply the Theorem 1 by distinguishing three cases,[2] which are outlined in Fig. 1: (1) Let us suppose that the nature plays H, Frances observes H, and $\gamma_E\gamma_F + v_H(1+\gamma_F) \leq 1$. From her point of view, if Emil plays $x \leq (1-v_H)/(1+\gamma_F)$, then she plays $y_H \geq (v_H + \gamma_F)/(1+\gamma_F)$ satisfying $x + y_H = 1$. Thus, the outcome game is agreement, otherwise, the negotiation breaks down. (2) Let us suppose that the nature state is L, Frances observes L, and $\gamma_E\gamma_F \leq 1$. From her point of view, if Emil plays $x \leq 1/(1+\gamma_F)$, then she plays $y_L \geq \gamma_F/(1+\gamma_F)$ satisfying $x + y_L = 1$. Thus, the outcome game is agreement, otherwise, the negotiation breaks down. (3) Let us suppose that $\gamma_E\gamma_F + \theta v_H(1+\gamma_E) \leq 1$ for

[2] We eliminate the overline $\overline{x}$, $\overline{y}_H$ and $\overline{y}_L$ writing x, y_H, y_L for simplifying the notation.

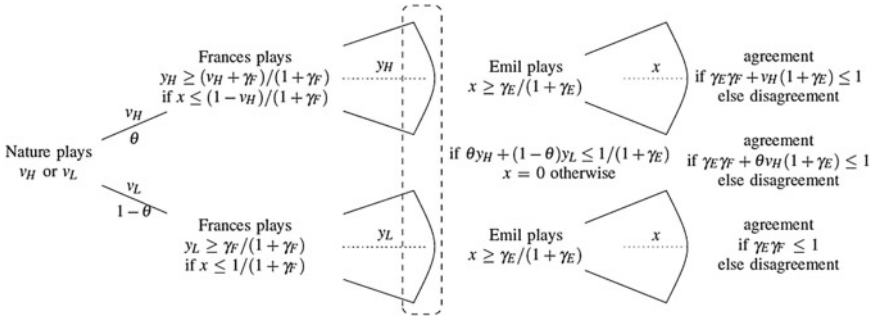


Fig. 1 Asymmetric information. Frances observes her outside opportunity and the Emil's outside opportunity completely. Instead Emil does not observe the Frances's outside opportunity, but he observes his outside opportunity completely. They play simultaneously

any nature state, which Emil does not observe. From Emil's point of view, if Frances plays y_H, y_L such that $\theta y_H+(1-\theta)y_L \leq 1/(1+\gamma_E)$, then he plays $x \geq \gamma_E/(1+\gamma_E)$ satisfying $x+\theta y_H+(1-\theta)y_L \leq 1$. Thus, the outcome game is agreement, otherwise, the negotiation breaks down.

In summary, the behavior of the players and the outcome of the game heavily depends on inequality $\gamma_E\gamma_F + \theta v_H(1+\gamma_E) \leq 1$. (1) When $\gamma_E\gamma_F > 1$, i.e. there is not a probability θ satisfying the inequality, then the achievement of an agreement is not possible. (2) When $\gamma_E\gamma_F \leq 1$, then there exists a number $\hat{\theta} \geq 0$ (not necessarily ≤ 1) such that $\gamma_E\gamma_F + \hat{\theta} v_H(1+\gamma_E) = 1$, thus for all probability $0 \leq \theta \leq \max\{\hat{\theta}, 1\}$ the agreement is obtained only if the nature plays L and disagreement otherwise. (3) Let us observe that, if $\gamma_E\gamma_F \leq 1$ and $\hat{\theta} > 1$, then an agreement is guarantied independent of what the nature plays.

Example 1 Let us suppose that $\gamma_E \doteq \gamma_F \doteq 3/4$. The random variable v takes the values $v_H \doteq 3/4$ and $v_L \doteq 0$ with probability θ and $1-\theta$ respectively. (1) Let us suppose that nature state is H. From Frances's point of view, as $\gamma_E\gamma_F+v_H(1+\gamma_E) = 15/8 > 1$, negotiation breaks down. Thus, Frances obtains $y_H = v_H = 3/4$ and Emil obtains $x = 0$. (2) Let us suppose that the nature state is L. From Frances's point of view, as $\gamma_E\gamma_F = 9/16 \leq 1$, she plays $y_L \geq \gamma_F/(1+\gamma_F) = 3/7$, if Emil plays $x \leq 1/(1+\gamma_F) = 4/7$ satisfying $x + y_L = 1$. The outcome is agreement. (3) The parameter $\hat{\theta} = (1-\gamma_E\gamma_F)/(v_H(1+\gamma_E)) = 1/3$, then for all $\theta \leq \hat{\theta} = 1/3$, the outcome game is agreement if the nature plays L and, otherwise, disagreement. A particular case is $\theta \doteq 1/4 \leq \hat{\theta} = 1/3$, thus from Emil's point of view, as $\gamma_E\gamma_F + \theta v_H(1+\gamma_F) = 57/64 \leq 1$, he plays $x \geq \gamma_E/(1+\gamma_E) = 3/7$ if Frances plays y_H, y_L satisfying $\theta y_H + (1-\theta)y_L \leq 1/(1+\gamma_E) = 4/7$ and $x + \theta y_H + (1-\theta)y_L = 1$. The outcome is agreement if the nature plays L (with probability $1 - \theta = 3/4$) and, otherwise, disagreement.

Symmetric and Complete Information Let us suppose that Frances's outside opportunity v is a random variable that takes two possible values $0 < v_H \leq 1$ (high) and

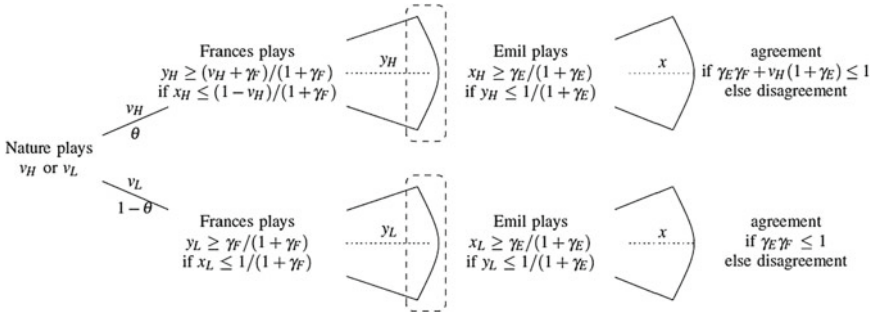


Fig. 2 Symmetric and complete information. Frances and Emil observe their outside opportunity and the outside opportunity of the other player completely. Both players play simultaneously

0 (low) with probability θ and $1-\theta$ respectively (with $\theta \in]0,1[$). Emil's outside opportunity is $u \doteq 0$ and the pie is $\pi \doteq 1$. Both players have complete information of all game parameters.

Emil chooses strategies x_H, x_L, and y_H, y_L respectively for Frances, depending if the nature plays H or L. The agents try to maximize the function utility given by $f(x_H, x_L, y_H, y_L) \doteq \theta(x_H - \gamma_E y_H) + (1-\theta)(x_L - \gamma_E y_L)$ and $g(x_H, x_L, y_H, y_L) \doteq \theta(y_H - \gamma_F x_H) + (1-\theta)(y_L - \gamma_F x_L)$ for Emil and Frances respectively, where $\gamma_E, \gamma_F \geq 0$ represent the externalities (Fig. 2).

In this case we say that the strategy tuple $(\overline{x}_H, \overline{x}_L, \overline{y}_H, \overline{y}_L)$ is a Nash equilibrium of the game if and only if

$$f(\overline{x}_H, \overline{x}_L, \overline{y}_H, \overline{y}_L) = \max_{x_H, x_L} \{f(x_H, x_L, \overline{y}_H, \overline{y}_L) : x_H, x_L \geq 0, x_H + \overline{y}_H \leq 1, x_L + \overline{y}_L \leq 1, \\ x_L - \gamma_E \overline{y}_L \leq 0 \text{ and } x_H - \gamma_E \overline{y}_H \geq 0\};$$

and

$$g(\overline{x}_H, \overline{x}_L, \overline{y}_H, \overline{y}_L) = \max_{y_H, y_L} \{g(\overline{x}_H, \overline{x}_L, y_H, y_L) : y_H, y_L \geq 0, \overline{x}_H + y_H \leq 1, \overline{x}_L + y_L \leq 1, \\ y_L - \gamma_F \overline{x}_L \leq 0 \text{ and } y_H - \gamma_F \overline{x}_H \geq v_H\};$$

To solve this problem, we apply Theorem 1 by distinguishing four cases: (1) The nature plays H, Emil and Frances observe H, and $\gamma_E\gamma_F + v_H(1+\gamma_F) \leq 1$. From Frances's point of view, she plays $y_H \geq (v_H + \gamma_F)/(1+\gamma_F)$, if Emil plays $x_H \leq (1-v_H)/(1+\gamma_F)$. The outcome game is agreement satisfying $x_H + y_H = 1$, and disagreement otherwise. (2) The nature plays L, Emil and Frances observe L, and $\gamma_E\gamma_F \leq 1$. From Frances's point of view, she plays $y_L \geq \gamma_F/(1+\gamma_F)$ if Emil plays $x_L \leq 1/(1+\gamma_F)$ satisfying $x_L + y_L = 1$. The game outcome is agreement and, otherwise, disagreement. (3) The nature plays H, Emil and Frances observe H, and $\gamma_E\gamma_F + v_H(1+\gamma_E) \leq 1$. From Emil's point of view, he plays $x_H \geq \gamma_E/(1+\gamma_E)$, if Frances plays $y_H \leq 1/(1+\gamma_E)$, satisfying $x_H + y_H = 1$. The game outcome is agreement and, otherwise, disagreement. (4) The nature plays L, Emil and Frances

observe L, and $\gamma_E\gamma_F \leq 1$. From Emil's point of view, he plays $x_L \geq \gamma_E/(1+\gamma_E)$, if Frances plays $y_L \leq 1/(1+\gamma_E)$, satisfying $x_L + y_L = 1$. The game outcome is agreement, and disagreement otherwise.

Example 2 Let us suppose that $\gamma_E \doteq \gamma_F \doteq 1/2$. The random variable v takes the values $v_H \doteq 3/4$ and $v_L \doteq 0$ with probability θ and $1-\theta$ respectively. (1) The nature plays H, Emil and Frances observe H, and $\gamma_E\gamma_F + v_H(1+\gamma_F) = 11/8 \geq 1$. Therefore, the negotiation breaks, thus Frances obtains $y_H = v_H = 3/4$ and Emil obtains $x = 0$. (2) The nature plays L, Emil and Frances observe L, and $\gamma_E\gamma_F = 1/4 \leq 1$. From Frances's point of view, she plays $y_L \geq \gamma_F/(1+\gamma_F) = 1/3$ if Emil plays $x_L \leq 1/(1+\gamma_F) = 2/3$ satisfying $x_L + y_L = 1$. The game outcome is agreement and disagreement otherwise. (3) The nature plays L, Emil and Frances observe L, and $\gamma_E\gamma_F \leq 1$. From Emil's point of view, he plays $x_L \geq \gamma_E/(1+\gamma_E) = 1/3$, if Frances plays $y_L \leq 1/(1+\gamma_E) = 2/3$, satisfying $x_L + y_L = 1$. The game outcome is agreement and disagreement otherwise. (4) Because (2) and (3), the game outcome is agreement.

3.2 Dynamic Bargaining Under Uncertainty

Now we extend the model developed in the last section into the dynamic case. Let us suppose that the Frances's outside opportunity are random variable v_t, which follows a stochastic process taking in each period $t \in \{0, 1, \ldots\}$ the value $0 < v_H \leq 1$ (high) or $v_L = 0$ (low). The initial probability is θ_0 of the value v_H and $1-\theta_0$ of v_L. In each period, this probability is updated according to the following Markov process $p(v_t|v_{t-1})$ given by the matrix

$$\begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix} \quad \text{where } \alpha, \beta \in]0, 1[.$$

From the Markov process theory, we know that the marginal probability $p(v_t)$ for the period t is given by the product

$$p(v_t) = p(v_0)p(v_t|v_{t-1})^t = \begin{pmatrix} \theta_0 & 1-\theta_0 \end{pmatrix} \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}^t \quad \text{for } t \in \{0, 1, \ldots\}.$$

The square matrix $p(v_t|v_{t-1})$ can be computed by using the product QDQ^{-1}, where D is the 2×2-diagonal matrix of the eigenvalues of $p(v_t|v_{t-1})$ and Q is the 2×2-matrix, where the columns are the corresponding independent eigenvectors. Therefore, the power matrix $p(v_t|v_{t-1})^t$ is given by QD^tQ^{-1}. In this case, the eigenvalues are $\alpha + \beta - 1$ and 1 and the corresponding eigenvectors are the columns $(1, (1-\beta)/(1-\alpha))$ and $(1, 1)$ respectively.

Thus, we can compute the marginal probability $p(v_t)$, by using the expression QD^tQ^{-1}. Therefore, $p(v_t)$ is given by

Table 1 Behavior patterns

Behavior pattern	Condition
1 Possible agreement	$\theta_0 \leq \theta_\infty \leq \hat{\theta}$ or $\theta_\infty \leq \theta_0 \leq \hat{\theta}$
2 Only disagreement	$\hat{\theta} \leq \theta_0 \leq \theta_\infty$ or $\hat{\theta} \leq \theta_\infty \leq \theta_0$
3 Possible agreement $\rightarrow$ only disagreement	$\theta_0 \leq \hat{\theta} \leq \theta_\infty$
4 only disagreement (*delay*) $\rightarrow$ possible agreement	$\theta_\infty \leq \hat{\theta} \leq \theta_0$

The game has different outcome patterns depending heavily on the inequality $\gamma_E\gamma_F + \theta v_H(1+\gamma_E) \leq 1$. If $\hat{\theta}$ is such that $\gamma_E\gamma_F + \hat{\theta}v_H(1+\gamma_E) = 1$, θ_0 is the start probability, and $\theta_\infty \doteq \lim_{t\to\infty}\theta_t$ of evolution of θ_t, then, for all t such that $\theta_t \leq \max\{\hat{\theta}, 1\}$, the outcome game is an agreement and disagreement otherwise

$$\begin{pmatrix} \theta_t \\ 1-\theta_t \end{pmatrix} = \frac{1}{2-(\alpha+\beta)} \begin{pmatrix} (\alpha+\beta-1)^t((2-(\alpha+\beta))\theta+\beta-1)+1-\beta \\ (\alpha+\beta-1)^t(-(2-(\alpha+\beta))\theta-\beta+1)+1-\alpha \end{pmatrix}.$$

Furthermore, we can obtain the stationary probability when $t \to \infty$, which is independent of the initial state, is given by[3]

$$\theta_\infty \doteq \lim_{t\to\infty}\theta_t = \frac{1-\beta}{2-(\alpha+\beta)} \quad \text{and} \quad 1-\theta_\infty = \frac{1-\alpha}{2-(\alpha+\beta)},$$

which there exists if $\alpha + \beta \geq 1$.

Now, let us consider the sequence of probabilities $\{\theta_t\}$ and the case when there exists a number $\hat{\theta}$ that satisfies the equality $\gamma_E\gamma_F + \hat{\theta}v_H(1+\gamma_E) = 1$. Applying the results of Sect. 3.1, for all period t that $\theta_t > \hat{\theta}$, the game breaks down independently of the nature state in each period (*only disagreement* pattern). Otherwise, for all period t such that $\theta_t \leq \hat{\theta}$, the game breaks down if the nature plays H and it reaches an agreement otherwise (*possible agreement* pattern). To sump up, depending on the initial state θ_0 and on the evolution of θ_t [given by the Eq. (3.2)], the players can have several behavior patterns, which are showed on Table 1.

References

1. Chowdhury, P. R. (1998). Externalities and bargaining disagreement. *Economics Letters*, *61*(1), 61–65.
2. Jéhiel, P., & Moldovanu, B. (1995). Cyclical delay in bargaining with externalities. *The Review of Economic Studies*, *62*(4), 619–637.
3. Jéhiel, P., & Moldovanu, B. (1995). Negative externalities may cause delay in negotiation. *Econometrica*, *63*(3), 1321–1335.
4. Laengle, S., & Loyola, G. (2012). Bargaining and negative externalities. *Optimization Letters*, *6*(3), 421–430.
5. Nash, J. (1953). Two-person cooperative games. *Econometrica*, *21*(1), 128–140.

[3] The Markov process theory provides another form to obtain the stationary probability, by considering the row vector of probabilities q such that $qp(v_t|v_{t-1}) = q$, which coincides with the results presented here.

Marginal Cost of Capacity for the Case of Overlapping Capacity Investments

Christian Lohmann

Abstract We examine a setting where the owner of a company delegates the authority to make overlapping capacity investments to an impatient manager. If the manager's internal interest rate exceeds the owner's cost of capital, a discrepancy arises between the owner's and the manager's perceived marginal cost of capacity, which is based on future cash flows associated with new capacity investments. This, however, leads the manager to capacity underinvestment. We argue that by using the performance measure residual income, in conjunction with particular depreciation rules, such as the relative practical capacity (RPC) depreciation rule, it is possible to avoid creating an underinvestment incentive for the manager. We begin by examining the effect direction of a deviation from the RPC depreciation rule on the manager's perceived marginal cost of capacity, which is based on future cost charges associated with new capacity investments. We then analyze the magnitude of the distortion of the manager's perceived marginal cost of capacity if the most convenient straight-line depreciation rule and the annuity depreciation rule are used.

1 Introduction

The present analysis of performance measures that prompt managers to make capacity investment decisions is based on the framework of overlapping capacity investments that was analytically examined for the first time by [1]. On the basis of specific assumptions, [1] calculated the cost for one unit of capacity made available for one period of time, even though capacity investment expenditures are commonly incurred for capacity that can be used over multiple periods. References [2, 3] recently analyzed the relationship between the marginal and historical cost of capacity when

C. Lohmann (✉)
Schumpeter School of Business and Economics, University of Wuppertal,
42119 Wuppertal, Germany
e-mail: lohmann@wiwi.uni-wuppertal.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_38,

there is a sequence of overlapping capacity investments. Their studies show that the marginal cost of capacity corresponds to the average historical cost of capacity if particular depreciation rules, like the relative practical capacity (RPC) depreciation rule, are used.

In the following, we analyze managerial performance measures for this type of overlapping capacity investments. According to the scenario that we consider here, the owner of a company delegates the authority to make capacity investments to the manager. This scenario is realistic, provided that the manager has superior information about future demand on capacity and future attainable revenues from that capacity. The calculus of the investment decision follows the equation that marginal revenue of capacity is equal to marginal cost of capacity. If the manager has the same time preferences as the owner, then the manager will have appropriate capacity investment incentives to make optimal investment decisions by calculating the marginal cost of capacity on the basis of future cash flows associated with new capacity investments.

In our setting, we consider an impatient manager whose internal discount rate exceeds the owner's cost of capital. Due to the discrepancy in the time preferences, the manager's perceived marginal cost of capacity exceeds the owner's marginal cost of capacity when both marginal costs are calculated on the basis of future cash flows associated with new capacity investments. In that case, a serious underinvestment problem arises because the manager's perceived marginal cost of capacity exceeds that of the owner. However, if the manager is paid a constant share in the obtained residual income in each period, the owner may have the opportunity to determine the depreciation rule so that the manager's perceived marginal cost of capacity, which is based on future cost charges associated with new capacity investments, coincides with the owner's marginal cost of capacity.

The analysis of the RPC depreciation rule indicates that the manager's perceived marginal cost of capacity (which, as explained, is based on future cost charges associated with new capacity investments) coincides with the owner's marginal cost of capacity so the manager is offered the desired investment incentives. In contrast to that, the straight-line depreciation schedule can induce overinvestment or underinvestment incentives due to higher or lower manager's perceived marginal cost of capacity. In view of that, the objective of our paper is to quantify the magnitude of distortions in the manager's perceived marginal cost of capacity that are caused by the performance measure residual income in conjunction with the straight-line depreciation schedule.

2 Marginal Cost of Capacity Based on Future Cash Flows Associated with New Capacity Investments

The marginal cost of capacity $c(r)$ based on future cash flows associated with new capacity investments is given by Eq. (1) if the no-excess capacity condition holds (see [2, 3]).

$$c(r) = \frac{v}{\sum_{t=1}^{T} \frac{x_t}{(1+r)^t}} \tag{1}$$

A new investment in one unit of capacity requires the investment expenditure v and increases the available capacity x_t for the following periods $1 \leq t \leq T$ of the entire useful life of the asset T. For the special case of linear decay in capacity, x_t is given by $x_t = 1 - \beta \cdot (t-1)$, where the parameter β indicates the decay of one period and $\beta = 0$ corresponds to the one-hoss-shay setting. The marginal cost of capacity $c(r)$ is constant in time, decreasing in the useful lifetime T and increasing in the cost of capital r.

We now turn to a firm that is managed by an impatient manager. The manager's performance measure is cash flow and he or she is paid a constant share in the cash flow achieved in each period. The impatient manager calculates the marginal cost of capacity from his or her point of view on the basis of the internal interest rate r_M, which exceeds the owner's cost of capital r. According to Eq. (1), the manager perceives a higher marginal cost of capacity ($c(r_M) > c(r)$) as a result of his or her internal interest rate ($r_M > r$). That result indicates underinvestment behavior on the part of the manager. That result is constant in time as the deviation between the owner's marginal cost of capacity $c(r)$ and the manager's perceived marginal cost of capacity $c(r_M)$ is also constant in time.

In the following, we focus on the deviation in the marginal cost of capacity $c(r_M) - c(r)$ to estimate the magnitude of the underinvestment problem. For that purpose, we have to calculate the percentage deviation Δc.

$$\Delta c = \frac{c(r_M) - c(r)}{c(r)} \tag{2}$$

Figure 1 shows the percentage deviation Δc for a specific setting with linear decay in capacity $x_t = 1 - \beta \cdot (t-1)$. If $r_M = r$, the percentage deviation is given by $\Delta c = 0\,\%$. The percentage deviation Δc escalates rapidly if the manager's internal interest rate increases. In the common one-hoss-shay setting with $\beta = 0.0$, if the manager's internal interest rate is $r_M = 0.2$, this leads to a percentage deviation of about $\Delta c = 47\,\%$ and may cause a significant underinvestment problem. Furthermore, we can also see that increasing (linear) decay in capacity decreases the percentage deviation Δc. In contrast to that, the manager has an incentive to overinvest if his or her internal interest rate drops below the owner's cost of capital and the manager's performance measure is cash flow.

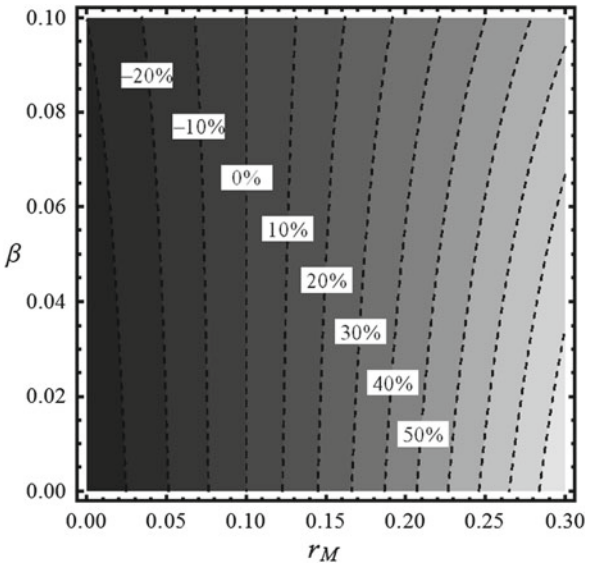


Fig. 1 Level sets of the percentage deviation Δc for different combinations of the manager's internal discount rate r_M and linear decay in productive capacity β. The parameters for this numerical simulation are: cost of capital $r = 0.1$ and useful life of assets $T = 10$ periods

3 Manager's Perceived Marginal Cost of Capacity Based on Future Cost Charges Associated with New Capacity Investments

The manager's perceived marginal cost of capacity $c(r_M, d)$ is determined by the cost charges $z_t(r, d)$ associated with a new investment in one unit of capacity. The cost charges $z_t(r, d)$ consist of depreciation and interest charges. The depreciation schedule d is described by the vector $d = (d_0, d_1, \ldots, d_T)$ with the property $\sum_{t=0}^{T} d_t = 1$, where d_t denotes the depreciation charge in period t after investment as a share in the investment expenditure $v = BV_0$. The book value at the end of period t is given by $BV_t = v \cdot (1 - \sum_{t=0}^{t} d_t)$. Note that $d_0 > 0$ reflects partial direct expensing, while $d_0 < 0$ can be interpreted as an initial write-up of the capacity asset. Thus, each cost charge can be calculated by $z_t(r, d) = d_t \cdot BV_0 + r \cdot BV_{t-1}(d)$ with $BV_{-1} = 0$, where $z_t(r, d) = d_0 \cdot BV_0$ corresponds to the partial direct expense charge for the investment expenditure v. If the manager's performance measure is residual income, the manager's perceived marginal cost of capacity $c(r_M, d)$ is given by Eq. (3).

$$c(r_M, d) = \frac{\sum_{t=0}^{T} \frac{z_t(r,d)}{(1+r_M)^t}}{\sum_{t=1}^{T} \frac{x_t}{(1+r_M)^t}} \tag{3}$$

If the manager's internal interest rate coincides with the owner's cost of capital ($r_M = r$), we can prove that $\sum_{t=0}^{T} \frac{z_t(r,d)}{(1+r)^t} = \sum_{t=0}^{T} \frac{d_t \cdot BV_0 + r \cdot BV_{t-1}(d)}{(1+r)^t} = v$ and therefore $c(r_M, d) = c(r)$. This equality is a direct consequence of the fundamental accounting identity, which shows that discounted future cash flows coincide with the sum of current book value and discounted future residual incomes (see [4, 5]). Furthermore, previous research has shown that there exists a depreciation schedule d^* for which the manager's perceived marginal cost of capacity $c(r_M, d^*)$, which is based on cost charges associated with new capacity investments, coincides with the owner's marginal cost of capacity $c(r)$. The depreciation schedule d^* is the result of the RPC depreciation rule (see [2, 3]).

If the manager's performance measure is residual income, the manager's perceived marginal cost of capacity $c(r_M, d)$ depends on the depreciation schedule d. If the depreciation schedule d is more accelerated or more decelerated than the RPC depreciation schedule d^*, the manager's perceived marginal cost of capacity $c(r_M, d)$ does not coincide with the marginal cost of capacity $c(r)$.

Straight-line depreciation schedules are most common in financial accounting and differ in the fraction $d_0^{sl} \geq 0$ of the investment that is expensed immediately. The straight-line depreciation schedule is described by the vector $d = (d_0^{sl}, d_1^{sl}, \ldots, d_T^{sl})$ with the properties $\sum_{t=0}^{T} d_t^{sl} = 1$ and $d_t^{sl} = \frac{1-d_0^{sl}}{T}$ for all $1 \leq t \leq T$. In the following, we analyze the straight-line depreciation schedule d^{sl} with $d_0^{sl} = 0$. Under the assumption that the capacity x_t linearly decays over time and is given by $x_t = 1 - \beta \cdot (t-1)$, [6] shows that the straight-line depreciation schedule d^{sl} coincides with the RPC depreciation schedule d^* if $d_0^{sl} = d_0^* = 0$ and the linear decay parameter β is equal to $\beta^* = \frac{r}{1+r \cdot T}$.

Lemma 1 *The straight-line depreciation schedule d^{sl} is more accelerated, or decelerated, than the RPC depreciation schedule d^* if $\beta < \beta^*$ or $\beta > \beta^*$ respectively.*

Proposition 1 *If the straight-line depreciation schedule d^{sl} is more accelerated than the RPC depreciation schedule d^*, then*

$$c(r_M, d^{sl}) = \begin{cases} \geq c(r) & \text{if } r_M \geq r \\ \leq c(r) & \text{if } r_M \leq r \end{cases}.$$

If the straight-line depreciation schedule d^{sl} is more decelerated than the RPC depreciation schedule d^*, then

$$c(r_M, d^{sl}) = \begin{cases} \leq c(r) & \text{if } r_M \geq r \\ \geq c(r) & \text{if } r_M \leq r \end{cases}.$$

The magnitude of the underinvestment and overinvestment incentives depends on the deviation in the marginal cost of capacity $c(r_M, d^{sl}) - c(r)$. Consequently, we have to calculate the percentage deviation $\Delta c(d^{sl})$ of the manager's perceived marginal cost of capacity $c(r_M, d^{sl})$ relative to the owner's marginal cost of capacity $c(r)$.

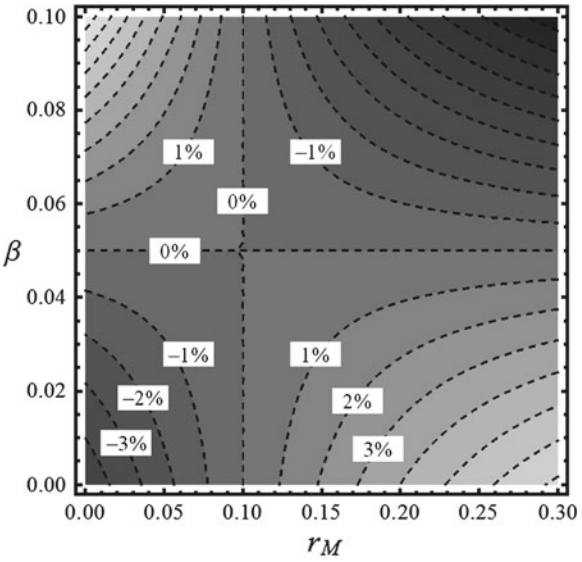


Fig. 2 Level sets of the percentage deviation $\Delta c(d^{sl})$ for different combinations of the manager's internal discount rate r_M and linear decay in productive capacity β ($\beta^* = 0.05$). The parameters for this numerical simulation are: cost of capital $r = 0.1$ and useful life of assets $T = 10$ periods

$$\Delta c(d^{sl}) = \frac{c(r_M, d^{sl}) - c(r)}{c(r)} \tag{4}$$

Figure 2 shows the percentage deviation $\Delta c(d^{sl})$ of the straight-line depreciation schedule for different combinations of the manager's internal discount rate r_M and linear decay in productive capacity β. For the given parameter values, the percentage deviations are remarkably low. That means that the straight-line depreciation schedule itself does not have a large impact on the manager's perceived marginal cost of capacity. Figure 2 suggests that the percentage deviation $\Delta c(d^{sl})$ is zero if $r_M = r$ or $\beta = \beta^*$. In the case of $r_M = r$, any depreciation schedule does not affect the manager's perceived marginal cost of capacity according to Proposition 1, and in the case $\beta = \beta^*$, the straight-line depreciation schedule corresponds to the RPC depreciation schedule according to Lemma 1. For the percentage deviations to increase by more than 5 % it would take an extremely impatient (or patient) manager and extreme decay parameters.

The main contribution of this paper to the literature on managerial incentives is that it analyzes the common straight-line depreciation schedules and its effect on the manager's perceived marginal cost of capacity, as well as their ability to achieve a periodic performance measure in terms of residual income, which provides an impatient manager with the desired investment incentives. The analysis shows that the structure of the depreciation schedule during the useful lifetime of an asset does not seem to be crucial to the periodic residual income performance measure for the

analyzed overlapping capacity investment setting. In particular, the distortions in the manager's perceived marginal cost of capacity, which follows from the performance measure residual income, remain small for a wide range of parameter constellations.

References

1. Arrow, K. (1964). Optimal capital policy, cost of capital and myopic decision rules. *Annals of the Institute of Statistical Mathematics*, *1–2*, 21–30.
2. Rogerson, W. (2008). Inter-temporal cost allocation and investment decisions. *Journal of Political Economy*, *116*, 931–950.
3. Rajan, M. V., & Reichelstein, S. (2009). Depreciation rules and the relation between marginal and historical cost. *Journal of Accounting Research*, *47*, 823–865.
4. Preinreich, G. (1938). Annual survey of economic theory: The theory of depreciation. *Econometrica*, *6*, 219–231.
5. Lücke, W. (1955). Investitionsrechnung auf der Basis von Ausgaben oder Kosten? *Z. handelswissenschaft. Forsch.*, *7*, 310–324.
6. Friedl, G. (2007). Ursachen und Lösung des Unterinvestitionsproblems bei einer kostenbasierten Preisregulierung. *Betriebswirt.*, *67*, 335–348.

Supply Chain Coordination Under Demand Uncertainty: Analysis of General Continuous Quantity Discounts

Hamid Mashreghi and Mohammad Reza Amin-Naseri

Abstract Quantity discount (QD) contracts are commonplace in theory and practice with different complexities. There exist two classes of QD contracts: continuous versus discrete schemes. Under continuous QD schemes, a general differentiable wholesale-price function is considered as a decreasing function of ordering quantities while in discrete QD schemes there is a price list with decreasing wholesale-price levels such as all-unit QD or incremental QD contracts. In this paper, we aim to analyze the structure of general continuous quantity discounts to coordinate a two-tier supply chain with additive demand uncertainty. We demonstrate sufficient coordination conditions based on joint-optimization of ordering and pricing decisions. Considering prerequisites for achieving coordination, the specific case of linear QD contract is analyzed and some applicable non-linear continuous QD schemes are introduced. Moreover application of such non-linear schemes is discussed for implementation of QDs (continues and discrete schemes) in real cases.

1 Introduction and Preliminary Review

Quantity discount (QD) contracts are frequently analyzed and implemented in theory and practice for supply chain (SC) management. QDs can be assumed as continuous or discrete schemes providing smaller wholesale-prices for larger ordering sizes. In the case of discrete QDs, literature mainly focuses on All-unit (AQD) or Incremental (IQD) schemes [4] while for the case of continuous QDs, the majority of the literature concentrate on Linear QDs (LQD) [2].

H. Mashreghi · M. R. Amin-Naseri (✉)
Department of Industrial Enginnering, Tarbiat Modares University, Tehran, Iran
e-mail: amin_nas@modares.ac.ir

H. Mashreghi
Cetim, LIACS, Leiden University, Niels Bohrweg 1, Leiden, The Netherlands
e-mail: mashreghy@gmail.com; mashreghi@modares.ac.ir

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_39,

Herein we assume a general continuous QD (GCQD) scheme for analyzing the possibility of SC coordination. Under GCQD the wholesale-price is defined as a general continuous decreasing function with respect to ordering quantity. Such comprehensive QD schemes are useful from both sides of analysis and practice. Analyzing a GCQD not only can provide better understanding about the familier continuous QDs such as LQD, but also enriches the design process of novel non-linear QD schemes for achieving coordination. Moreover, it can be useful for redesigning discrete QD contracts with diferent discrete levels instead of analytical continuous QD functions. In these cases, decision making for practitioners will become easier when a unique rule can be defined all over the price spectrum.

2 The Profit Maximization Model

Assume a two-tier SC with a GCQD contract i.e. the supplier charges $w_d = w_d(q)$ for every unit sold to the retailer as a wholesale-price which has a decreasing function in ordering quantity, q. Assume c_r and c_s as the retailer's and the supplier's unit marginal costs. Similarly consider g_r and g_s for the retailer's and the supplier's unit goodwill penalty costs when shortages occur. In addition if leftover would occur, the overages has disposal cost h and its negative measures where $c_r < h < 0$ can be interpreted as salvage value. Demand has additive uncertainty i.e. $D(\varepsilon, p) = a - bp + \varepsilon$ where $a, b > 0$ and p is the selling price. The term ε as the additive random part has pdf, $f(.)$, and cdf, $F(.)$. The stocking decision is defined as $z = q - y(p)$ where $y(p) = a - bp$ and w_d is rearranged in z and p as $w_d = w_d(z, p) = w_d(z + y(p))$.

Define $\dot{w}_z = \frac{\partial w_d(z,p)}{\partial z}$, $\dot{w}_p = \frac{\partial w_d(z,p)}{\partial p}$, $\ddot{w}_{zz} = \frac{\partial^2 w_d(z,p)}{\partial z^2}$, $\ddot{w}_{pp} = \frac{\partial^2 w_d(z,p)}{\partial p^2}$, $\ddot{w}_{zp} = \frac{\partial^2 w_d(z,p)}{\partial z \partial p}$, $\ddot{w}_{pz} = \frac{\partial^2 w_d(z,p)}{\partial p \partial z}$, and $\Delta_w = \ddot{w}_{zz}\ddot{w}_{pp} - \ddot{w}_{zp}\ddot{w}_{pz}$ where Δ_w is the determinant of Hesian matrix of $w_d(z, p)$. Proposition 1 demonstrates the general properties for any continuous QD under additive demand uncertainty.

Proposition 1 *Under any GCQD contract with additive demand uncertainty:*

1. $\dot{w}_z = \dot{w}_q$ *and therefore* $\dot{w}_z < 0$.
2. $\dot{w}_p > 0$. *Thus* $w_d(z, p) = f(p)$ *i.e. the wholesale-price is a function of selling price and therefore a price transmission from SC downstream is occured.*
3. $\ddot{w}_{zz} = \ddot{w}_{qq}$, $\ddot{w}_{pp} = b^2\ddot{w}_{qq} = b^2\ddot{w}_{zz}$, *and therefore* $\ddot{w}_{pp}$ *and* $\ddot{w}_{zz}$ *have similar sign.*
4. $\ddot{w}_{zp} = \ddot{w}_{pz} = -b\ddot{w}_{qq}$.
5. $\Delta_w = b^2\ddot{w}_{zz}^2 - \ddot{w}_{zp}^2$.

Proof Since $q = a - bp + z$, $\frac{\partial q}{\partial z} = 1$ and $\dot{w}_z = \dot{w}_q$ due to $\frac{\partial w}{\partial z} = \frac{\partial w}{\partial q} \cdot \frac{\partial q}{\partial z}$. Thus as $\dot{w}_q < 0$ we can show the statement (1). In addition, regarding $\frac{\partial w}{\partial p} = \frac{\partial w}{\partial p} \cdot \frac{\partial q}{\partial p}$, we have $\dot{w}_p = -b\dot{w}_q = -b\dot{w}_z$ because $\frac{\partial q}{\partial p} = -b$ and since $\dot{w}_q < 0$ we have $\dot{w}_p > 0$ which results in part (2). Moreover statements of (3)–(5) are proved based on the direct definitions of second derivatives, $\dot{w}_p = -b\dot{w}_q > 0$, and $\dot{w}_z = \dot{w}_q$. □

Defining $\Lambda(z) = \int_A^z (z-u) f(u) du$ and $\Theta(z) = \int_z^B (u-z) f(u) du$, the retailer's expected profit function become $E[(\pi_r(z, p)] = p[y(p) + \mu] - h\Lambda(z) - (p + g_r)\Theta(z) - (w_d(z, p) + c_r)[y(p) + z]$. Similarly, the expected profit function for the supplier and the SC become $E[(\pi_s(z, p)] = (w_d(z, p) + c_s)[y(p) + z] - g_s)\Theta(z)$ and $E[(\pi_S C(z, p)] = (p - c)[y(p) + \mu] - (c + h)\Lambda(z) - (p + g - c)\Theta(z)$.

Assuming demand cases with non-decreasing hazard rates for their probability density functions [1], the SC's expected profit function is strictly concave in z and p and the centralized optimal solution can be determined based on the first optimality conditions which results in $p^0 = p_c^0 - \frac{\Theta(z)}{2b}$ and $F(z^0) = \frac{p+g-c}{p+g+h}$ where p_c^0 is the SC's riskless optimal pricing decision.

3 General Conditions to Achieve Coordination

In order to achieve coordination with price-dependent demand, joint-optimization should be considered. Seeking the optimal decisions we solve simultaneously two first-order optimality conditions for the retailer. In addition, it is important to know whether these optimality conditions forces the retailer to stop ordering process. Morevoer, if such conditions allow the retailer to order more than zero, it should be asked that do these condisions allow the supplier to achieve nonzero profit? By answering to the first question the possibility of having a regular SC can be proved and by answering to the second question the occurrence of *Double Marginalization* [3] problem can be verified. Thus, checking $q > 0$ and $w_d > c_s$ provides the possibility of having optimal ordering or pricing decisions in order to achieve coordination.

3.1 Analysis of the Retailer's Optimality Conditions

The retailer's first optimality conditions are as follows:

$$\frac{\partial E[(\pi_r(z, p)]}{\partial z} = -hF(z) + (p + g_r)[1 - F(z)] - (w_d + c_r) - \dot{w}_z[y(p) + z] = 0 \quad (1)$$

$$\frac{\partial E[(\pi_r(z, p)]}{\partial p} = a + \mu - 2bp - \Theta(z) + b(w_d + c_r) - \dot{w}_p[y(p) + z] = 0. \quad (2)$$

3.1.1 Coordination by Aligning Optimal Stocking Decisions

Assume z^0 satisfies the condition (1). Rearranging the condition with $g_r = g - g_s$ follows that $\frac{\partial E[(\pi_r(z^0, p)]}{\partial z} = g^s \cdot \frac{p+g-c}{p+g+h} + (c_r - w - g_r) - \dot{w}_z[y(p) + z^0] = 0$ which leads to $w = c_s - \frac{g_s(h+c)}{p+g+h} - \dot{w}_z[y(p) + z^0]$. As $\dot{w}_z < 0$ for having a profitable supplier

$(w > c_s)$, the condition $-\frac{g_s(h+c)}{\dot{w}_z(p+g+h)} < q^0$ has to be satisfied by SC's ordering decisions where $q^0 = y(p^0) + z^0 \geq 0$. It shows that having profitable supplier to prevent double marginalization is obtained by setting optimal ordering decision (or equivalent stocking decision) larger than this parametric threshold $-\frac{g_s(h+c)}{\dot{w}_z(p+g+h)}$.

3.1.2 Coordination by Aligning Optimal Pricing Decisions

The optimal pricing can coordinate SC if p^0 satisfies the condition (2) follows that $\frac{\partial E[(\pi_r(z,p^0)]}{\partial p} = a+\mu-2bp_c^0+b(w_d+c_r)-\dot{w}_p q(p^0) = 0$. Since $a+\mu-2bp_c^0 = -bc$, $bw = bc_s+\dot{w}_p q(p^0)$ and consequently $w = c_s+\frac{\dot{w}_p q(p^0)}{b}$. Thus if $q^0 = y(p^0)+z^0 \geq 0$ we have $w > c_s$ because $\dot{w}_p > 0$ and $b > 0$. It shows that ordering by the retailer can guarantee having profitable supplier and coordination can be achieved by aligning optimal pricing decisions.

3.1.3 Coordination by Joint Optimization of Ordering and Pricing

Previously, coordination was achieved separately by ordering and pricing decisions if $w - c_s = \alpha(p^0, z) = \frac{\dot{w}_p}{b}[y(p^0) + z]$ and $w - c_s = \beta(p, z^0) = -\frac{g_s(h+c)}{p+g+h} - \dot{w}_z[y(p)+z^0]$. With joint optimal pair of stocking and pricing, (p^0, z^0), $y(p^0)+z = y(p)+z^0 = y(p^0)+z^0 = q^0$. Moreover, since $\dot{w}_p = -b\dot{w}_z, \alpha(p^0, z^0) = \beta(p^0, z^0)$. Therefore, seeking the necessary condition to achieve coordination by simultaneous optimization results in $\frac{g_s(h+c)}{p^0+g+h} = 0$. Since $h > -c_r$, $h > -c$ and $h + c > 0$. Consequently SC coordination by joint optimization is obtained by $g_s = 0$. Therefore under coordinating GCQD contract, the whole responsibility of underage costs is assigned to the retailer to prevent non-profitable supplier and double marginalization.

3.2 Analysis of the Supplier's Optimality Conditions

The supplier's first optimality conditions are as follows:

$$\frac{\partial E[(\pi_s(z, p)]}{\partial z} = \dot{w}_z[y(p) + z] + (w_d - c_s) + g_s[1 - F(z)] = 0 \tag{3}$$

$$\frac{\partial E[(\pi_s(z, p)]}{\partial p} = \dot{w}_p[y(p) + z] - b(w_d + c_s) = 0 \tag{4}$$

The condition (3) shows that $q = y(p)+z = -\frac{g_s(1-F(z))+(w_d-c_s)}{\dot{w}_z}$. Since $\dot{w}_z < 0$ we have $q > 0$ only if $(w_d - c_s) > -g_s(1 - F(z))$. Therefore it is reasonable to have $(w_d - c_s) < 0$ where $q > 0$ and the optimal ordering cannot prevent

Table 1 Some possible non-linear continues QD schemes

Name	$w(q)$	$w(z, p)$	$\dot{w}_z < 0$
Power-n	$w - dq^n$	$w - d(z + y(p))^n$	$-nd(z + y(p))^{n-1}$
Adjusted power-n	$w - dq - y(p)^n$	$w - dz^n$	$-ndz^{n-1}$
Exponential	$w - de^q$	$w - de^{z+y(p))}$	$-de^{z+y(p))}$
Logarithmic	$w - d\text{Ln}(q)$	$w - d\text{Ln}(z + y(p))$	$\frac{-d}{z+y(p)}$

double marginalization. However, (4) shows that $(w_d - c_s) = \frac{\dot{w}_p[y(p)+z]}{b}$. Since $\dot{w}_p = -b\dot{w}_z > 0$ and $w_d = f(p)$, $q = \frac{b(w_d-c_s)}{\dot{w}_p} = -\frac{b(w_d-c_s)}{\dot{w}_z} > 0$ if $w_d > c_s$. Thus the optimal pricing can prevent double marginalization. Moreover considering the results under $\dot{w}_p = -b\dot{w}_z$ follows that $\frac{g_s(1-F(z))}{\dot{w}_z} = 0$. It shows that the joint optimization is possible when $g_s = 0$ or $\dot{w}_z \to 0$ which means GCQD contract changes into a price-only contract which contradicts the assumptions.

4 Analysis of Linear and Non-linear QDs

The majority of the literature concentrates on LQDs as a continuous QD scheme. With a LQD scheme we have $w_d(q) = w - dq = w_d(z) - dy(p)$ where w and d are positive constant parameters. The term d (the QD's slope in q) increases the retailer's expected profit and decreases the supplier's expected profit functions. Assuming LQD as a GCQD it can be seen that $w(z, p) = w - ad + bdp - dz$, $\dot{w}_z = -d, \dot{w}_p = bd > 0$, and $\ddot{w}_{zz} = \ddot{w}_{pp} = \ddot{w}_{zp} = \ddot{w}_{pz} = \Delta_w = 0$. As $\dot{w}_p = -b\dot{w}_z$, the coordination can be achieved only by $g_s = 0$.

Moreover, regardless of assuming linear functions we can choose other kinds of continuous schemes as volume discounts for the retailers and the end customers (Table 1). For instance, with n-degree NCQD scheme ($n > 0$) the amount of discount grows faster in comparison to the linear fashion. With non-linear functions, the main assumption that should be considered is $\dot{w}_z < 0$. In addition, larger n makes larger discount rate. Similarly, the exponential schemes can be interpreted as a poly-nominal discount using tailor-series. On the other hand, with a logarithmic QD scheme the seller wants to decrease the rate of discounting in comparison to the linear, polynomial or exponential cases. Thus, we can use other qualities for a GCQD contract under the main condition, $\frac{\dot{w}_p}{\dot{w}_z} = -b$.

5 Concluding Remarks and Further Research

According to the coordination analysis it can be seen that achieving coordination is possible by aligning the retailer's ordering or pricing optimal decisions with the centralized decisions. In order to achieve coordination by joint optimization the SC's

goodwill penalty cost should be compensated by the retailer i.e. $g = g_r$. For the the supplier's decisions, achieving coordination is merely possible by aligning optimal pricing decisions, or aligning optimal ordering decisions with $g_s = 0$. This condition is sufficient to achieve coordination by joint optimization of pricing and ordering decisions. Therefore, the scenario $g_s = 0$ can provide a negotiable parameter-setting for the SC partners toward achieving coordination by ordering, pricing or joint optimization from both sides of the retailer and the supplier.

Although the literature focuses on achieving coordination by synchronizing ordering decisions and assumes the supplier as an price-taker, it seems that the possibility of achieving coordination by pricing decision or joint optimization can provide better collaborative space to motivate and involve the supplier in pricing and marketing decisions. Such assumption introduces the supplier as an effective SC's agent in price-setting process, provides reinforced partnership with long-term relationships, and promotes SC's brand in the market. For this reason, these contractual mechanisms should be analyzed by further numerical studies.

Moreover, the analysis of LQD contract as well-known continuous QD contracs can be used to achieve coordination based on the introduced scenarios of GCQD scheme. Moreover, introducing some possible non-linear continuous QD schemes compared to the case of LQDs shows that they can be implemented to achieve coordination for a two-tier SC where using them would be easier by adjusting discrete levels of price lists for real practitioners. Finally, for future research, it is necessary to develop and analyze non-linear QD schemes based on their practical interpretations in actual cases.

References

1. Petruzzi, N., & Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research*, *47*, 183–194.
2. Qin, Y., Wang, R., Vakharia, A. J., Chen, Y., & Seref, M. M. H. (2011). The newsvendor problem: Review and directions for future research. *European Journal of Operational Research*, *213*, 361–374.
3. Spengler, J. (1950). Vertical integration and antitrust policy. *Journal of Political Economy*, *58*(4), 347–352.
4. Weng, Z. K. (1995). Channel coordination and quantity discounts. *Management Science*, *41*(9), 1509–1522.

An Integer Programming Model for the Hospitals/Residents Problem with Couples

Iain McBride and David F. Manlove

Abstract The Hospitals/Residents problem with Couples (HRC) is a generalisation of the classical Hospitals/Residents problem (HR) that is important in practical applications because it models the case where couples submit joint preference lists over pairs of (typically geographically close) hospitals. In this paper we give a new NP-completeness result for the problem of deciding whether a stable matching exists, in highly restricted instances of HRC. Further, we present an Integer Programming (IP) model for HRC and extend it the case where preference lists can include ties. Further, we describe an empirical study of an IP model for HRC and its extension to the case where preference lists can include ties. This model was applied to randomly generated instances and also real-world instances arising from previous matching runs of the Scottish Foundation Allocation Scheme, used to allocate junior doctors to hospitals in Scotland.

1 Introduction

The National Resident Matching Program (NRMP) matches graduating medical students to hospitals in the US, matching 25,526 students in 2012. Similarly, in Scotland, until recently, medical graduates were matched to Foundation Programme places via the Scottish Foundation Allocation Scheme (SFAS). Centralised matching schemes such as NRMP and SFAS have had to evolve to accommodate linked couples who wish to be allocated to (geographically) compatible hospitals. The requirement to consider the joint preferences of couples has been in place in the NRMP context

I. McBride (✉) · D. F. Manlove
School of Computing Science, University of Glasgow, Sir Alwyn Williams Building, Glasgow G12 8QQ, UK
e-mail: i.mcbride.1@research.gla.ac.uk

D. F. Manlove
e-mail: David.Manlove@glasgow.ac.uk

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_40,

since 1983 and more recently in the case of SFAS. The underlying allocation problem for NRMP and SFAS can modelled by the so called Hospitals/Residents Problem with Couples (HRC).

An instance of the *Hospitals Residents Problem with Couples* consists of a set of *hospitals* H and a set of *residents* R. The residents in R are partitioned into two sets, S and S'. The set S consists of *single* residents and the set S' consists of those residents involved in *couples*. There is a set $C = \{(r_i, r_j): r_i, r_j \in S'\}$ of *couples* such that each resident in S' belongs to exactly one pair in C.

Each single resident $r_i \in S$ expresses a linear preference order over some subset of the hospitals in H, representing the hospitals that resident r_i finds *acceptable*; any hospital not in this subset is therefore *unacceptable* to r_i. Each pair of residents $(r_i, r_j) \in C$ expresses a joint linear preference order over a subset A of $H \times H$ where $(h_p, h_q) \in A$ represents the joint assignment of r_i to h_p and r_j to h_q. The hospital pairs in A represent those joint assignments that are acceptable to (r_i, r_j), all other joint assignments being unacceptable to (r_i, r_j).

Each hospital $h_j \in H$ expresses a linear preference order over those residents who find h_j acceptable, either as a single resident or as part of a couple. Also, each hospital $h_j \in H$ has a *capacity*, c_j, its maximum number of available posts.

The preferences expressed in this fashion are reciprocal: if a resident r_i is acceptable to a hospital h_j, either as a single resident or as part of a couple, then h_j is also acceptable to r_i, and vice versa. A many-to-one *matching* between residents and hospitals is sought, which is a set of acceptable resident-hospital pairs such that each resident appears in at most one pair and each hospital appears in a number of pairs that does not exceed its capacity. Further, each couple (r_i, r_j) is either jointly unmatched, meaning that both r_i and r_j are unmatched, or jointly matched to some pair (h_k, h_l) that (r_i, r_j) find acceptable.

In an HRC instance we seek a *stable* matching, which guarantees that no resident and hospital, and no couple and pair of hospitals, has an incentive to deviate from their assignments and become matched to each other.

Roth [8] considered stability in the HRC context although did not define the concept explicitly. However, a variety of stability definitions do exist in the HRC context [2, 3, 5]. The definition of stability applied in the work which follows is that given by McDermid and Manlove in [5], shown below in Definition 1, which gives those mutually acceptable pairs, (r_i, h_k) and $((r_i, r_j), (h_k, h_l))$, whose existence would block a matching in HRC.

Definition 1 A matching M is stable if none of the following holds:

1. The matching is blocked by a hospital h_j and a single resident r_i, as in the classical HR problem.
2. The matching is blocked by a couple (r_i, r_j) and a hospital h_k such that *either*
   (a) (r_i, r_j) prefers $(h_k, M(r_j))$ to $(M(r_i), M(r_j))$, and h_k is either undersubscribed in M or prefers r_i to some member of $M(h_k)\backslash\{r_j\}$ *or*
   (b) (r_i, r_j) prefers $(M(r_i), h_k)$ to $(M(r_i), M(r_j))$, and h_k is either undersubscribed in M or prefers r_j to some member of $M(h_k)\backslash\{r_i\}$

3. The matching is blocked by a couple (r_i, r_j) and (not necessarily distinct) hospitals $h_k \neq M(r_i)$, $h_l \neq M(r_j)$; that is, (r_i, r_j) prefers the joint assignment (h_k, h_l) to $(M(r_i), M(r_j))$, and *either*
   (a) $h_k \neq h_l$, and h_k (respectively h_l) is either undersubscribed in M or prefers r_i (respectively r_j) to at least one of its assigned residents in M; *or*
   (b) $h_k = h_l$, and h_k has at least two free posts in M, i.e., $c_k - |M(h_k)| \geq 2$; *or*
   (c) $h_k = h_l$, and h_k has one free post in M, i.e., $c_k - |M(h_k)| = 1$, and h_k prefers at least one of r_i, r_j to some member of $M(h_k)$; *or*
   (d) $h_k = h_l$, h_k is full in M, h_k prefers r_i to some $r_s \in M(h_k)$, and h_k prefers r_j to some $r_t \in M(h_k) \backslash \{r_s\}$.

An instance of HRC need not admit a stable matching [9]. Also an instance may admit stable matchings of differing sizes [1]. Further, the problem of deciding whether there exists a stable matching in an instance of HRC is NP-complete, even in the restricted case where there are no single residents and all of the hospitals have only one available post [6, 7].

Let (α, β)-HRC denote the restriction of HRC in which each single resident's preference list contains at most α hospitals, each couple's preference list contains at most α pairs of hospitals and each hospital's preference list contains at most β residents. In many practical applications the residents' preference lists are short. However, the problem remains hard even in this case and Manlove and McDermid [5] showed that $(3, 6)$-HRC is NP-complete.

In Sect. 2 of this paper we present a new NP-completeness result for the problem of deciding whether there exists a stable matching in an instance of $(2, 3)$-HRC and a summary of an Integer Programming (IP) model for finding a maximum cardinality stable matching in an instance of HRC. Further, in Sect. 3 we present an empirical study of this model as applied to randomly generated instances and also real-world instances arising from previous matching runs of SFAS. Some conclusions are given in Sect. 4.

2 Complexity of HRC and IP Model

In a technical report by the same authors [4] we prove the following new result; for space reasons the details of the proof are omitted.

Theorem 1 *Given an instance of* $(2, 3)$-HRC*, the problem of deciding whether the instance supports a stable matching is NP-complete. The result holds even if there are no single residents and each hospital has capacity 1.*

In [4] we give an IP model for finding a maximum cardinality stable matching in HRC. Each model has $O(m)$ binary-valued variables and $O(m + cL^2)$ constraints where m is the total length of the hospitals' preference lists, c is number of couples and L is the maximum length of a couple's preference list. The space complexity

of each model is $O(m(m + cL^2))$ and each model can be built in $O(m^4)$ time. For space reasons the details of the models are omitted.

3 Empirical Results

We ran experiments on a Java implementation of the IP model as described in [4] applied to both randomly-generated and real data. We present data showing (1) the average time taken to find a maximum cardinality stable matching or report that no stable matching exists, and (2) the average size of a maximum cardinality stable matching where a stable matching did exist. All experiments were carried out on a desktop PC with an Intel i5-2400 3.1 GHz processor, with 8 Gb of memory running Windows 7. The IP solver used in all cases was CPLEX 12.4 and the model was implemented in Java using CPLEX Concert.

To test our implementation for correctness we used a brute force algorithm which recursively generated all possible matchings admitted by an HRC instance and selected a maximum cardinality stable matching from amongst those matchings or reported that none of the generated matchings was stable. Due to the inefficiency of this algorithm it may only be realistically applied to relatively small instances. When solving several thousand HRC instances involving up to 15 residents our implementation agreed with the brute force algorithm when reporting whether the instance admitted a stable solution and further our implementation returned a stable matching of the same size as a maximum cardinality stable matching output by the brute force algorithm.

Experiments with randomly generated instances—In our first experiment, we report on data obtained as we increased the number of residents while maintaining a constant ratio of couples, hospitals and posts to residents. For various values of x ($100 \leq x \leq 1{,}000$) in increments of 30, 1,000 randomly generated instances were created containing x residents, $0.1x$ couples and $0.1x$ hospitals with x available posts which were unevenly distributed amongst the hospitals.

The data in Fig. 1 show that the mean time to find a maximum cardinality stable matching increased as we increased the number of residents in the instance. Figure 1 also shows that the percentage of HRC instances that admit a stable matching does not appear to be correlated with the number of residents involved in the instance and that as the number of residents in the instances increased, the mean size of the maximum cardinality stable matching supported by the instances increased linearly with the number of residents involved the instance. In the second experiment, we report data as we increased the the percentage of residents involved in couples while maintaining the same total number of residents, hospitals and posts. For various values of x ($0 \leq x \leq 250$) in increments of 25, 1,000 randomly generated instances were created containing 1000 residents, x couples (and hence $1{,}000 - 2x$ single residents) and 100 hospitals with 1,000 available posts which were unevenly distributed amongst the hospitals.

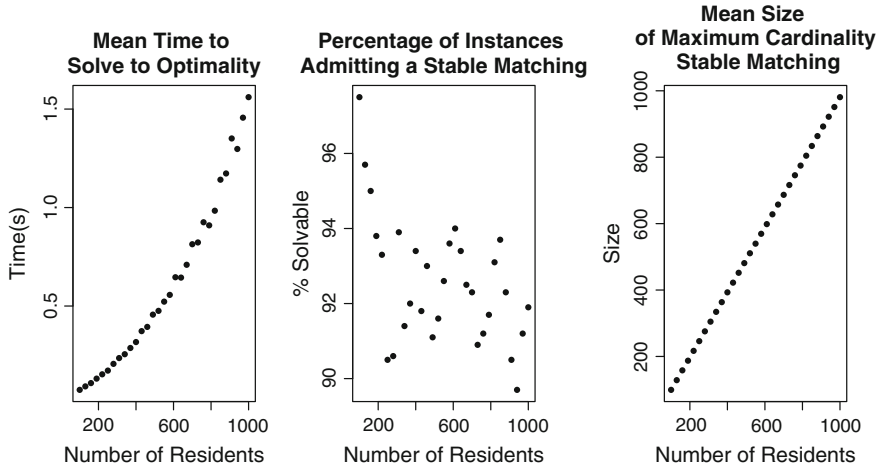


Fig. 1 Data obtained when attempting to find a maximum cardinality stable matching in randomly generated instances from experiment 1

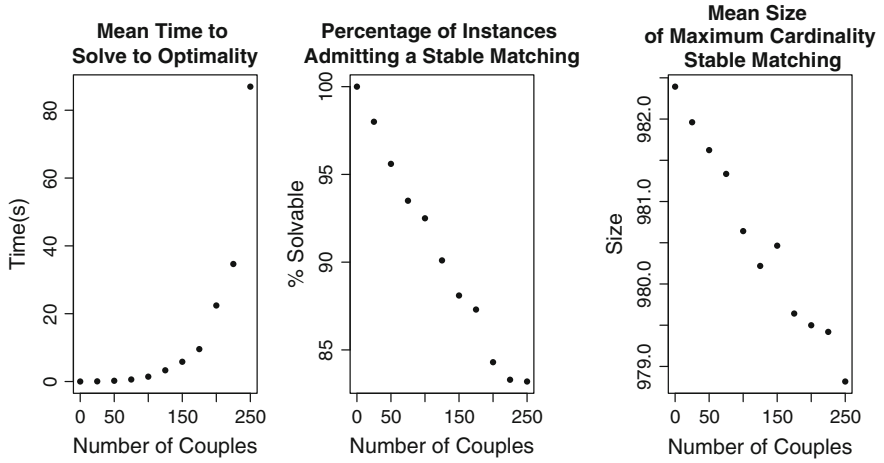


Fig. 2 Data obtained when attempting to find a maximum cardinality stable matching in randomly generated instances from experiment 2

The data in Fig. 2 show that the mean time to find a maximum cardinality stable matching increased as we increased the number of residents involved in couples. Further, Fig. 2 shows that the percentage of HRC instances admitting a stable matching fell as the percentage of residents in the instances involved in couples increased. When 50 % of the residents in the instance were involved in a couple, 832 of the 1,000 instances admitted a stable matching. Figure 2 also shows that as the percentage of residents in the instances involved in couples increased the mean size of a maximum cardinality stable matching tended to decrease.

Table 1 Results obtained from the previous 3 years' SFAS data

	Number of residents	Number of couples	Number of hospitals	Number of posts	Max cardinality stable matching	Time to solution (s)
2012	710	17	52	720	681	9.62
2011	736	12	52	736	688	10.41
2010	734	20	52	735	681	33.92

Performance of the model with real world data—The *Hospitals/Residents Problem with Couples and Ties* (HRCT) is a generalisation of HRC in which hospitals (respectively residents) may find some subsets of their acceptable residents (respectively hospitals) equally preferable. Residents (respectively hospitals) that are found equally preferable by a hospital (respectively resident) are *tied* with each other in the preference list of that hospital (respectively resident). It is straightforward to adapt Definition 1 to the HRCT case.

SFAS assigned junior doctors to two-year training posts in Scotland. In this process the hospitals' preferences were derived from the residents' *scores*, where a junior doctor's score was derived from their previous academic performance. If two residents received the same score, they were tied in a hospital's preference list. Thus, the underlying SFAS matching problem may be correctly modelled by HRCT.

Hence, we further extended our implementation to solve instances of HRCT as described in [4] and were able to find a maximum cardinality stable matching admitted by the real data obtained from the SFAS context. The sizes of the maximum cardinality stable matchings obtained in the SFAS context for the 3 years to 2012 are shown in Table 1 alongside the time taken to find these matchings.

4 Conclusions

We conclude that the IP model presented in this paper performs well when finding a maximum cardinality stable matching in instances that are similar to those that arose from the SFAS application. It remains to investigate the performance of the model as we increase the size of the instance beyond that of the SFAS application.

Acknowledgments Iain McBride: Supported by a SICSA Prize PhD Studentship. David F. Manlove: Supported by Engineering and Physical Sciences Research Council grant GR/EP/K010042/1.

References

1. Aldershof, B., & Carducci, O. M. (1996). Stable matching with couples. *Discrete Applied Mathematics*, *68*, 203–207.
2. Biró, P., Irving, R. W., & Schlotter, I. (2011). Stable matching with couples: An empirical study. *ACM Journal of Experimental Algorithmics*, *16*, Section 1, article 2, 27 p.

3. Gusfield, D., & Irving, R. W. (1989). *The stable marriage problem: Structure and algorithms*. Cambridge: MIT Press.
4. McBride, I., & Manlove, D. F. (2013). *The hospitals/residents problem with couples: Complexity and integer programming models*. Technical Report, Computing Research Repository, Cornell University Library.
5. McDermid, E. J., & Manlove, D. F. (2010). Keeping partners together: Algorithmic results for the hospitals/residents problem with couples. *Journal of Combinatorial Optimization*, *19*(3), 279–303.
6. Ng, C., & Hirschberg, D. S. (1991). Three-dimensional stable matching problems. *SIAM Journal on Discrete Mathematics*, *4*, 245–252.
7. Ronn, E. (1990). NP-complete stable matching problems. *Journal of Algorithms*, *11*, 285–304.
8. Roth, A. E. (1984). The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy*, *92*(6), 991–1016.
9. Roth, A. E. (1990). New physicians: A natural experiment in market organization. *Science*, *250*, 1524–1528.

Techno-economic Analysis and Evaluation of Recycling Measures for Iron and Steel Slags

Christoph Meyer, Matthias G. Wichmann and Thomas S. Spengler

Abstract Iron and steel production involves the generation of numerous different by-products. An essential group of by-products are iron and steel slags which can be recycled to obtain secondary resources. In order to decide how slags are to be recycled, a large number of technical, economic and ecological variables has to be considered. An approach to recycling planning taking all relevant variables into account is not known. This contribution introduces a recycling planning approach for slags based on a techno-economic analysis and evaluation of recycling measures.

1 Introduction

With a total of 1.5 billion tons of crude steel in 2012 worldwide steel production has reached its highest level to date. This requires to deal with large amounts of by-products. In Germany, the production of 42.7 million tons of crude steel led to 13.4 million tons of iron and steel slags in 2012. Slags perform important metallurgical tasks and are inevitable for iron and steel production processes. Although slag production is inevitable, slags are not considered waste and can be used as secondary resources, e.g. road construction material, cements and fertilizers.

For slag recycling there is a variety of alternative recycling measures strongly depending on a multitude of technical, economic and ecological variables. For exam-

C. Meyer (✉) · M. G. Wichmann · T. S. Spengler
Institute of Automotive Management and Industrial Production,
Technische Universität Braunschweig, Katharinenstr. 3,
D-38106 Braunschweig, Germany
e-mail: christoph.meyer@tu-bs.de

M. G. Wichmann
e-mail: ma.wichmann@tu-bs.de

T. S. Spengler
e-mail: t.spengler@tu-bs.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_41,

ple the chemical and mechanical composition of recycled slags is subject to laws and also influences the attainable product price. From the perspective of an iron and steel producer who is legally obliged to deal with accruing by-products this leads to the question how slags are to be recycled.

This contribution introduces an operative recycling planning approach for slags considering technical, economic and ecological variables. In Sect. 2 the planning task and problem characteristics are addressed in more detail. In Sect. 3 a model formulation incorporating a techno-economic analysis and evaluation of recycling measures is developed. In Sect. 4 the model is illustrated using a numerical example. The contribution closes with a conclusion and an outlook in Sect. 5.

2 Planning of Slag Recycling

Planning slag recycling needs to incorporate three dimensions. These are slag production, slag recycling and the usage of recycled slags as secondary resources. Slags are produced in different sources and vary in composition and amount, e.g. blast furnace slag or steelmaking slag. According to composition and amount slags can be recycled in different ways. Depending on how slags are recycled different sinks, e.g. road construction or cement production can be addressed.

The three dimensions can be regarded as elements of a network structure consisting of sources, recycling measures and sinks. For this network the aim of operative recycling planning is to determine the relevant material and energy flows from sources to recycling measures to sinks with regard to quantities and values. In order to determine the material and energy flows three categories of requirements have to be considered. These are technical, economic and ecological requirements.

From a technical perspective, slag recycling is based on mechanical and chemical process engineering [1]. Therefore the allocation of material and energy flows requires a sufficiently detailed representation of mechanical and chemical processes. In order to illustrate this, an example of a common recycling measure is given below. First, liquid slag is led into a slag pit where it continually solidifies. Second, the solid slag is processed through a configuration of crushers and screens. Depending on the production units and operating points technical parameters of the slag such as the grain size distribution can be altered. Varying with the grain size there are different applications for the resulting secondary resources, e.g. road construction material or coarse aggregates in concrete. Although existing approaches in recycling planning comprise descriptions of mechanical and chemical processes, e.g. the recycling of construction waste in [4] or electric arc furnace dusts in [2], the specifics of recycling iron and steel slags have not yet been considered.

From an economic perspective, slag recycling is characterized by recycling costs and revenues from sales of secondary resources. Determining recycling costs requires information on the relevant cost categories and their manifestation in a management accounting system. Determining revenues requires information on the target market for the secondary resources produced. The target market is characterized by regional

sales depending on the local supply of secondary resources and seasonality. Therefore the attainable price for secondary resources depends on the quantities to be sold and the current season. Examples of planning approaches incorporating these characteristics can be found in [4].

From an ecological perspective, slag recycling must comply with versatile regulations, e.g. legislation and standardization. Among others, these regulations specify the chemical composition of recycled slags. Examples of production planning approaches considering chemical composition limits can be found in [3] and [2].

As described above, existing approaches show congruencies with the mentioned requirements, e.g. concerning the representation of revenues for secondary resources or their chemical composition. Nevertheless there is no approach fulfilling all necessary requirements in technical, economic and ecological regard at the same time. In particular this applies to the incorporation of technical specifics of slag recycling into a planning approach. Therefore a mathematical model of the planning problem is formulated in the next section.

3 Model Formulation

The aim of recycling planning for slags is to determine the relevant material and energy flows in the recycling network with regard to quantities and values. Therefore the model formulation is based on a techno-economic analysis and evaluation of recycling measures.

Since the recycling measures for slags are primarily based on mechanical and chemical process engineering a sufficiently detailed but flexible process representation is needed for the quantity structure of the model. A flexible approach appropriate for the representation of such processes is activity analysis [4]. Activity analysis allows for modeling by-production of several fractions of secondary resources through crushing and screening as mentioned above. Besides, activity analysis can also be combined with approaches such as flowsheet simulation resulting in a detailed process representation [5]. Based on the process representation in the quantity structure a value structure comprising recycling costs and revenues of secondary resources is formulated.

Due to clarity reasons only a simplified model with emphasis on activity analysis and variable revenues is presented below. Therefore aspects such as the transportation of secondary resources are omitted from the formulation. The notation of the resulting model is as follows.

The material flows in the recycling network are represented as quantities of objects j, determined for a considered period t. Object j can refer to input as well as output objects. The transformation of input objects (slags) to output objects (secondary resources) is represented by recycling activities i. These recycling activities provide different operating points m. The objects representing secondary resources are allocated to compatible sinks k. On this basis the following model is derived.

$$\text{Max } CM = \sum_{t=1}^{T}\sum_{k=1}^{K} p_{k,t}\left(z_{k,t}\right)\cdot z_{k,t} - \sum_{t=1}^{T}\sum_{j=1}^{J} c_j^s \cdot s_{j,t} - \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{m=1}^{M} c_{i,m}^r \cdot \lambda_{i,m,t} \quad (1)$$

subject to

$$z_{k,t} = \sum_{j=1}^{J} z_{j,k,t} \quad \forall k,t \quad (2)$$

$$z_{j,k,t} \le a_{j,k} \cdot \left(s_{j,t-1} - s_{j,t} + y_{j,t}\right) \quad \forall j,k,t \quad (3)$$

$$s_{j,t} = s_{j,t-1} + y_{j,t} - \sum_{k=1}^{K} z_{j,k,t} \quad \forall j,t \quad (4)$$

$$s_{j,t} \le S_j^{\max} \quad \forall j,t \quad (5)$$

$$y_{j,t} = x_{j,t} + \sum_{i=1}^{I}\sum_{m=1}^{M} v_{i,j,m} \cdot \lambda_{i,m,t} \quad \forall j,t \quad (6)$$

$$\sum_{m=1}^{M} u_{i,m} \cdot \lambda_{i,m,t} \le U_i^{\max} \quad \forall i,t \quad (7)$$

$$z_{j,k,t}, s_{j,t}, \lambda_{i,m,t} \ge 0 \quad \forall i,j,k,m,t \quad (8)$$

$$s_{j,0} = 0 \quad \forall j \quad (9)$$

The objective function (1) aims at maximizing the contribution margin of the considered secondary resources. The contribution margin is the difference between revenues for secondary resources and cumulated inventory and recycling costs. The revenues are obtained by multiplying the price $p_{k,t}\left(z_{k,t}\right)$ with product quantity $z_{k,t}$ allocated to sink k in period t. Due to the mentioned market characteristics the revenue gained for a product depends on the product quantity allocated to a sink. The inventory costs for products are obtained by multiplying specific inventory costs c_j^s with the inventory quantity $s_{j,t}$ of object j in period t. The recycling costs are obtained by multiplying specific recycling costs $c_{i,m}^r$ with an activity level $\lambda_{i,m,t}$ for activity i in operating point m and period t. In addition to the objective function (1) the model incorporates constraints (2)–(9) which are explained below.

Constraint (2) describes the total product amount $z_{k,t}$ allocated to sink k in period t as sum of product amounts $z_{j,k,t}$, whereby $z_{j,k,t}$ represents the specific amount of object j allocated to sink k in period t. Hence, the overall demand $z_{k,t}$ of one sink k can be satisfied by multiple product amounts $z_{j,k,t}$.

Constraint (3) ensures that only compatible products are allocated to respective sinks. Here, the allocation of product amounts $z_{j,k,t}$ to a sink k is limited due to technical and ecological reasons using a binary parameter $a_{j,k}$. The parameter is multiplied with a large number $\left(s_{j,t-1} - s_{j,t} + y_{j,t}\right)$ which is composed of possible inventory changes $s_{j,t-1} - s_{j,t}$ and the object quantity $y_{j,t}$ produced or consumed

in the recycling network. Because of j representing input as well as output objects, $y_{j,t}$ represents quantities of input as well as output objects in period t.

Constraint (4) is the inventory equation. Thus, inventory quantities $s_{j,t}$ of object j in period t are equal to the inventory quantity in the previous period plus the production or consumption of object quantities $y_{j,t}$ minus the object quantities allocated to the entirety of sinks. Constraint (5) ensures that the maximum inventory capacity is not exceeded.

Constraint (6) describes the connection of the produced or consumed object quantities with an input parameter $x_{j,t}$ and the recycling activities. Therefore a recycling coefficient $v_{i,j,m}$ is multiplied with the activity level $\lambda_{i,m,t}$ thereby connecting the object quantities to the extent of activity usage.

Constraint (7) ensures that the extent of activity usage does not exceed a given capacity. Therefore the activity level $\lambda_{i,m,t}$ is multiplied with a utilization factor $u_{i,m}$.

Constraint (8) ensures that $z_{j,k,t}$, $s_{j,t}$ and $\lambda_{i,m,t}$ are nonnegative and Constraint (9) initializes the inventory quantities to be zero.

Apart from $p_{k,t}\left(z_{k,t}\right)$ not specified here, the model formulation yields a linear model and allows for a simple representation of the underlying recycling planning problem. Depending on the actual choice of function for $p_{k,t}\left(z_{k,t}\right)$ the model can become nonlinear.

4 Numerical Example

In order to illustrate the model a numerical example is given. The model was implemented in LINGO 14 and solved using a standard PC with 2.67 GHz and 4 GB RAM. The example network consists of one source, two recycling measures and three sinks, whereby two consecutive months are considered. The data used is oriented towards real quantities and values.

The source is given through a quantity of 90,000 t/month of blast furnace slag ($j = 1$) to be completely recycled. For recycling two activities can be used. Recycling activity 1 ($i = 1$) leads to the production of granulated blast furnace slag ($j = 2$) whereas recycling activity 2 ($i = 2$) produces air-cooled blast furnace slag in a coarse ($j = 3$) and a fine fraction ($j = 4$). Therefore activity 2 features two operating points. Using the first (second) operating point 60 % (40 %) coarse and 40 % (60 %) fine fraction are produced. According to the activity and the operating point used specific recycling costs $c^r_{i,m}$ ranging from 2 to 3 EUR/t are considered. The secondary resources produced are allocated exclusively to one sink representing potential buyers such as the cement industry. In order to incorporate an attainable price $p_{k,t}$ depending on the total product amount $z_{k,t}$ the linear function in Eq. (10) is considered, therefore yielding a quadratic objective function.

$$p_{k,t} = p^{\max}_{k,t} - b_{k,t} \cdot z_{k,t} \tag{10}$$

Table 1 Results of numerical example

Sink k ($t = 1$)	Price $p_{k,t}$	Quantity $z_{k,t}(t)$	Sink k ($t = 2$)	Price $p_{k,t}$	Quantity $z_{k,t}(t)$
1	2.29 EUR/t	14,152	1	4.04 EUR/t	26,101
2	1.58 EUR/t	42,326	2	2.33 EUR/t	41,521
3	2.09 EUR/t	27,081	3	2.84 EUR/t	28,817

$b_{k,t}$ is determined by maximum prices $p_{k,t}^{\max}$ ranging from 3 to 6 EUR/t ($z_{k,t} = 0$) and maximum product amounts $z_{k,t}$ ranging from 60,000 to 100,000 t/month ($p_{k,t} = 0$). In order to comprise a possible disposal of superfluous product quantities also negative prices are allowed. Apart from allocation to a sink, product quantities can also be stored in the first period for specific inventory costs of 0.5 EUR/(t·month). An excerpt of the results of this numerical example is given in Table 1.

Table 1 shows the allocated product amounts and the corresponding prices. Because of a simplified quantity structure that does not involve material losses it can be derived from the table that the input quantities of blast furnace slag are almost completely allocated to the three sinks in the first month. Merely a combined amount of 6,440 t of the three considered products is stored and allocated in the second month. This can be explained by slightly higher maximum prices assumed for the second month on account of seasonality. The solution of the numerical example leads to a contribution margin of 37,716 EUR.

5 Conclusion and Outlook

This contribution introduces a recycling planning approach for the recycling of iron and steel slags. The planning task and the problem characteristics are discussed. Based on this a mathematical formulation of the planning problem considering technical, economic and ecological variables is developed. In order to validate the model a numerical example is given.

Further research is necessary concerning a sufficiently detailed representation of the mechanical and chemical processes used for recycling iron and steel slags. In order to incorporate more complex process representations into the techno-economic analysis and evaluation, an approach based on flowsheet simulation is promising. Depending on the actual price function used in the mathematical formulation, the solubility of the model needs to be considered for practical problem sizes.

References

1. Das, B., Prakash, S., Reddy, P. S. R., & Misra, V. N. (2007). An overview of utilization of slag and sludge from steel industries. *Resources, Conservation and Recycling*, *2007*(50), 40–57.
2. Fröhling, M., Schwaderer, F., Bartusch, H., & Rentz, O. (2010). Integrated planning of transportation and recycling for multiple plants based on process simulation. *European Journal of Operational Research*, *207*(2), 958–970.
3. Klingelhöfer, H. E. (2000). *Betriebliche Entsorgung und Produktion*. Wiesbaden: DUV.
4. Spengler, T. S. (1994). *Industrielle Demontage-und Recyclingkonzepte*. Berlin: ESV.
5. Spengler, T. S., Hähre, S., Sieverdingbeck, A., & Rentz, O. (1998). Stoffflußbasierte Umweltkostenrechnung zur Bewertung industrieller Kreislaufwirtschaftskonzepte. *Zeitschrift für Betriebswirtschaft*, *1998*(2), 147–174.

Dynamical Supply Networks for Crisis and Disaster Relief: Networks Resilience and Decision Support in Uncertain Environments

Silja Meyer-Nieberg, Erik Kropat and Patrick Dolan Weber

Abstract Recent natural disasters affected many parts of the world and resulted in an extensive loss of life and disruption of infrastructure. The randomness of impacts and the urgency of response efforts require a rapid decision making in an often uncertain and complex environment. In particular, the organization and controlling of efficient humanitarian supply chains are challenging the operational analyst from both the theoretical and practical perspective. A far-sighted and comprehensive emergency planning can alleviate the effects of sudden-onset disasters and facilitate the efficient delivery of required commodities and humanitarian aid to the victims. Methods from computational networks and agent-based modelling supported by sophisticated data farming experiments allow a detailed analysis of network performance measures and an evaluation of the vulnerability of infrastructure and supply networks. These approaches can be used for relief planning as well as for a simulation of continuous aid work threatened by severe disruptions. This paper presents a first step towards an integrated dynamic network optimization approach which combines forecasting models and simulation.

1 Introduction

On January 12, 2010, a devastating earthquake struck Haiti near the capital Port-au-Prince. An estimated three million people were affected by the quake and approximately 250,000 people died. Camps for displaced people sprang up throughout

S. Meyer-Nieberg (✉) · E. Kropat
Universität der Bundeswehr München, Werner-Heisenberg Weg 39,
85577 Neubiberg, Germany
e-mail: silja.meyer-nieberg@unibw.de

E. Kropat
e-mail: erik.kropat@unibw.de

P. D. Weber
University of Arizona, 1401 E University Blvd, Tucson, AZ 85721, USA
e-mail: patrick310@email.arizona.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_42,

Port-au-Prince and other cities and at the peak 1.5 million people were living in refugee camps [2]. Within the first days after the disaster, the United Nations cluster coordination system was activated [1] and the international community launched a massive humanitarian response that was considered by the International Federation of Red Cross and Red Crescent Societies as "the largest humanitarian operation carried out in a single country" [5, p. 2]. Besides many other factors the logistical obstacles were immense: The Port-au-Prince seaport was damaged and non-functional and the international airport was operated initially with line-of-sight landings on one runway. The traffic infrastructure was severely damaged and hampered the response efforts considerably. The situation in the densely populated camps became even more complicated in October 2010, when a cholera outbreak threatened the health of refugees [6]. Though most cases of symptomatic cholera cases are considered as mild or moderate, an estimated 20 % of the total infections can cause severe dehydration from watery diarrhea that can kill within hours if left untreated [3].

This paper is a first step towards an investigation on how optimization coupled with forecasting and simulation models can contribute to the development of *dynamic humanitarian supply networks*. Humanitarian supply chains differ in many aspects considerably from the traditional commercial logistic chains [4]. For example, for-profit logistics tries to minimize transportation costs, usually based on stable and predictable demand patterns, whereas humanitarian supply chains aim at maximizing the demand satisfaction of the affected population while minimizing delivery times. The development of a distribution system is exacerbated by the fact that a dynamic change of parameters such as the state of the infrastructure (roads, airport, seaport), the availability and storage of goods, and the number of healthy and sick refugees in the camps has to be taken into account. In addition, several interconnected problems have to be addressed simultaneously, e.g., vehicle routing, truck assignment, demand forecasting, and epidemic modelling.

This paper describes a first approach towards an *integrated dynamic supply network optimization*. It provides the means to incorporate the results from forecasting and simulation in the process of supply network optimization in case of a major disaster. The corresponding *evolving computational supply networks* aim to maximize the assistance to the affected population and it explores how to ward off diseases by adapting the distribution patterns of the supply chain.

This paper is structured as follows. The next section introduces the scenario we considered. Afterwards the mathematical optimization model is described before presenting the results from some scenario calculations.

2 The Scenario

In the scenario under consideration, we address a *dynamical multi-commodity supply network* consisting of refugee camps, distribution centers, and main entry points to the country or region. The goal is to optimize the distribution of goods such as water, food, medicine and shelters with regard to the population in need, and the health status

of the people. As sometimes a huge number of people is living in relatively small area, the outbreak of diseases such as acute diarrhea and cholera has to considered. Such a situation changes the priority of goods to be delivered at the camps and has to be taken into account for the *dynamic distribution planning*. In our model, the *distribution policy* prioritizes camps where the gap between the demand and resource is the largest while it also reflects the current priority of the goods in the camp.

For the supply network, a homogeneous fleet of trucks is shipping the goods between the facilities. The actual input at the main entry points (airports, seaports, main roads) depends on the state of the infrastructure can increase with ongoing reconstruction efforts. For each type of good, standardized pallets are used for the transport between the sites, where each pallet can hold one type of good. Each facility is able to store goods up to a maximal storage capacity.

3 Relief Distribution

This section describes a first optimization model. The model will be coupled with forecasting and simulation models which give estimates for the population development inside the refugee camp and may model the outbreak of diseases. The aim is two-fold. On the one hand, it is possible to analyze the robustness of the supply chain plan by varying the model parameters. On the other hand, the simulation can be used for forecasting during relief operations. The model introduced here, represents a first step of the way. At present, we are interested in demonstrating the general feasibility of the approach.

The model considers a multi-stage approach with T time intervals (ti) of duration t_d. We consider the distribution of four goods: shelters (s), food (f), water (w), and medical supplies (m). The approach can of course be used for a general number of goods. Let $G := \{f, w, m, s\}$ be the set of goods under consideration. There are K refugee camps with a population $p_k(t)$ at time t. The demand of the camps depends on number of people inside the camp and the current state of the population. Healthy people require f_h amounts of food (kg/time interval), whereas sick people need f_s (kg/time interval). Similarly, the requirements for water read w_h (ℓ/time interval) for healthy and w_s (ℓ/time interval) sick persons. We also consider medical supplies with m_h (kg/time interval) and m_s (kg/time interval). In the case of shelters, we assume a maximal number of persons per shelter n_s. The requirements give raise to the demands $d_k^f(t)$, $d_k^w(t)$, $d_k^m(t)$, and $d_k^s(t)$.

Since we consider T time intervals, we have to derive estimates of future states, either by modelling using dynamical systems, agent-based simulation, or by applying forecast models.

The aim of the optimization is to meet the demand—especially if there are many sick people for instance during an epidemic. If an outbreak of a dangerous disease occurs, the priorities of fulfilling the single demands may change: In the case of a cholera outbreak, water and medical supplies (IV bags, salient solutions) may

become more important. To incorporate changing priorities in the model, we consider priority coefficients $\rho_k^f(t)$, $\rho_k^w(t)$, $\rho_k^m(t)$, and $\rho_k^s(t)$ which depend on the state of the population and can also be used to include the uncertainty of future states.

The refugee camps have a storage area with maximal capacities for the considered goods cap_{rk} with presently stored supplies $g_{rk}(t)$ for each good $g \in G$. This leads to the requirement

$$0 \leq \sum_{g \in G} g_{rk}(t) \leq cap_{rk} \tag{1}$$

for $t \in \{1, \ldots, T+1\}$. The shipping and distribution of the goods will be modeled by assuming standardized pallets of size. Each pallet takes one type of good with no_g units of good $g \in G$ per palette.

There are J distribution centers the location of which has been fixed beforehand. Similarly to the refugee camps, the distribution centers have maximal capacities cap_{dj} for the goods under consideration and stored supplies

$$0 \leq \sum_{g \in G} g_{dj}(t) \leq cap_{dj} \tag{2}$$

for $t \in \{1, \ldots, T+1\}$. The I primary points of entries for supplies are airports, seaports, and main roads. If there is a road, a storage area may be associated with it at a location outside the disaster area. All ports have a time dependent input $In_{pi}^g(t)$ for a good $g \in G$. and also a maximal storage capacity. The actual input depends on the state of the infrastructure at time t. We assume that the amount $In_{pi}^g(t)$ of good $g \in G$ can be distributed in the time interval t. Since the storage of supplies is given by $g_{pi}(t)$, $0 \leq \sum_{g \in G} g_{pi}(t) \leq cap_{pi}$ for $t \in \{1, \ldots, T+1\}$, the maximal amount that a primary entry point can provide is $In_{pi}^g(t) + g_{pi}(t)$ for all $g \in G$. We assume that the amount of incoming goods cannot be influenced. The model, however, can be easily adapted to this situation. We need to address the transport between the sites. We assume that it is possible to derive an estimate for the travel time $t_{lh}(t)$ between site l and site h. The estimate—for instance the mean travel time—could be based on data from the previous time intervals considering also newly arrived information on the state of the road network. We are considering a fleet of homogeneous vehicles (trucks) which is based at a site. In our first approach, we assume that each vehicle must return at the end of the interval and will travel just once. This leads to the following formulation. Goods can be transported between sites if $t_{lh}(t) \leq c_t t_d$ with t_d the length of the time interval and c_t a constant with $c_t \in]0, 0.5[$. In this case, the sites are directly connected. For the remainder of the paper, $\mathbb{1}_{\{t_{lh}(t) \leq c_d t_d\}}(h)$ is the corresponding indicator function.

Each site l can assign a maximal number of trucks N_l^{truck} for the transport. Each truck can be assigned exactly once for a transport to a site h leading to

$$\sum_h y_{lh}(t) \leq N_l^{\text{truck}} \tag{3}$$

with y_{lh} the decision variable counting the number of trucks assigned for the transport between l and h. Goods can only be transported between sites if two conditions are met. First, there must be trucks assigned for transport and the sites must be reachable in the time interval. First, the next series of decision variables is introduced with $x_{lh}^g(t)$ the number of pallets with good $g \in G$ transported at t between site l and site h. By introducing the constraint

$$x_{lh}^g(t) \;\leq\; \mathbb{1}_{\{t_{lh}(t) \leq c_d t_d\}}(h) \max_{pal} N_l^{\text{truck}} \quad \forall\, g \in G \tag{4}$$

with $\max_{pal}$ the maximal number of pallets a vehicle can load, transport between sites that are too far away is precluded, while

$$\sum_{g \in G} x_{lh}^g(t) \leq \max_{pal} y_{lh} \tag{5}$$

allows only to transport goods within the capacity of the trucks assigned. The amount of goods that can be assigned to shipments depends to the stored amount and the delivered amount

$$\sum_h x_{lh}^g(t) \leq g_{*l}(t-1) + \sum_k x_{kl}^g(t-1) \tag{6}$$

with $*$ standing for either a primary point of entry, a distribution center, or possibly a refugee camp if they are allowed to operate as distribution centers. The stored amount is then updated according to

$$g_{*l}(t) \leq g_{*l}(t-1) + \sum_k x_{kl}^g(t-1) - \sum_h x_{lh}^g(t). \tag{7}$$

In the case of the primary entry points, the equations read

$$\sum_h x_{ph}^g(t) \leq g_p(t-1) + In_p^g(t) \quad \forall\, g \in G \tag{8}$$

$$g_p(t) \leq g_p(t-1) + In_p^g(t) - \sum_h x_p^g(t) \quad \forall\, g \in G \tag{9}$$

For refugee camps k, define first the differences between current resources and the demand

$$\Delta_k^g(t) := g_{rk}(t) + \sum_l x_{lk}^g(t-1) - d_{kpal}^g(t) \quad \forall\, g \in G \tag{10}$$

with $d^g_{kpal}(t)$ denoting the number of required pallets of good g. It is the aim to improve the delivery of the important goods to the camps. Therefore, we consider the camps where the gap between demand and resource is largest—weighted by the current priority for the good in the camp. The optimization strives to make the gap as small as possible. One formulation for the objective function therefore reads

$$\max \sum_{t=1}^{T} \sum_{g \in G} \min_k \left\{ \rho^g_k(t) \mathbb{1}_{\{\Delta^g_k(t)<0\}} \left(\Delta^g_k(t)\right) \Delta^g_k(t) \right\}. \tag{11}$$

4 The Scenario

In a first basic scenario, two million people are evacuated to five refugee camps for the course of 14 days. There is a logistic growth of the camp population within 14 days and the population is equally divided among the five camps. We assume the following model for the required supply for person and time unit: (a) food: 3 units per person and day, (b) water: 1 unit per person and day, (c) shelters: 6 persons per shelter. The goods are transported by standard US trucks with a maximum load capacity of 24 pallets per truck. Each pallet can hold either 8,500 units of food, 750 units of water, 750 medical supplies or 4 shelters. In the basic scenario, we assume that there are unlimited storage capacities and that each site of the supply network can be reached by each other facility. There are two distribution centers and two ports. Each port can provide 150 shipments by trucks per time unit and each distribution center can ship at most 75 truck loads per time unit.

The GAMS solver with CPLEX needs 1,000 s on a Core i7 CPU with four 2.7 GHz-processing units to obtain a feasible solution with the quality of 14.5 % of the upper bound. The absolute value of the objective function reads 13,467. In total 65 solutions were obtained.

5 Conclusions and Outlook

This paper presented a first approach aimed at the *integrated dynamic supply network optimization*. A first model was obtained and solved for a specific evacuation scenario. In future studies, the present work will be extended combining the optimization with simulation and forecasting. We will investigate further optimization models, for instance vehicle routing, facility location and assignment problems under different types of uncertainty. To solve these tasks, heuristics and metaheuristics will be explored.

References

1. Bhattacharjee, A., & Lossio, R. (2011). *Evaluation of OCHA response to the Haiti earthquake*. Final report, January 2011. https://docs.unocha.org/sites/dms/Documents/Evaluation%20of%20OCHA%20Response%20to%20the%20Haiti%20Earthquake.pdf. Accessed 20 Aug 2013.
2. Bilham, R. (2010). Lessons from the Haiti earthquake. *Nature*, *463*, 878–979.
3. Butler, D. (2010). News Cholera tightens grip on Haiti. *Nature*, *468*, 483–484.
4. de la Torre, L. E., Dolinskaya, I. S., & Smilowitz, K. R. (2012). Disaster relief routing: Integrating research and practice. *Socio-Economic Planning Sciences*, *46*, 88–97.
5. International Federation of Red Cross and Red Crescent, Societies. *Haiti: Earthquake*. Operations, update no. 5, February 9, 2010.
6. Walton, D. A., & Ivers, L. C. (2011). Responding to cholera in post-earthquake Haiti. *The New England Journal of Medicine*, *364*(1), 3–5.

A Column Generation Approach to Home Care Staff Routing and Scheduling

Susumu Morito, Daiki Kishimoto, Hiroki Hayashi, Atsushi Torigoe, Shigeo Okamoto, Yuki Matsukawa and Nao Taniguchi

Abstract Daily route generation of home care staff is considered and a column generation heuristic is developed. Constraints considered include staff working hours, time window for each visit, means of transportation (bicycle/car), maximum allowable idle time between visits, patient/staff compatibility, among others. Since it is desired to generate compact routes in the geographically scattered area, the minimization of total travel time is used. Computational results based on real data will be presented. To further reduce CPU time, pre-processing of input data is performed to reduce the solution space by narrowing the time window of visits and by limiting candidate staff members who could be assigned to a specific visit. The pre-processing is performed by solving two small 0–1 programs. It is shown how the pre-processing cuts down the CPU time of the column generation algorithm.

1 Introduction

In home care medical services, nurses and physical therapists (we call them staff) visit homes of patients (we call them customers) and provide necessary medical services. Demand for home care services is increasing rapidly due to aging population. Currently, most of home care staff schedules are made manually. A feasible schedule is required to satisfy time windows for customer visits together with many other constraints, and its generation is difficult and takes long time. Needs for a computer-assisted planning system have increased for generating efficient routing and scheduling of home care staff.

S. Morito (✉) · D. Kishimoto · H. Hayashi · A. Torigoe
Waseda University, Shinjuku, Tokyo, Japan
e-mail: morito@waseda.jp

S. Okamoto · Y. Matsukawa · N. Taniguchi
Saint-Care Holding Corporation, Tokyo, Japan
e-mail: okamoto@saint-care.com

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_43,

Studies on routing and scheduling of home care medical as well as home helper services are recently performed in various countries. Eveborn et al. [1] is one of the early studies which formulated the problem as a set partitioning form and applied a heuristic algorithm. Therapist routing and scheduling has been studied by Shao et al. [4] and GRASP was used to solve the problem. Ikegami et al. [3] developed a min-cost flow-based efficient algorithm for home helper routing and scheduling, based on which a web-based scheduling system has been implemented.

2 Home Care Staff Routing and Scheduling

We seek a routing schedule of home care staff for a single day. The goal is to find routes for staff with minimum total travel time. Major assumptions are listed below:

1. A given set of customer visits must be all executed.
2. Service time and time window of each visit are given.
3. Each route starts and ends at the station.
4. Starting and ending times of a staff's work day are given.
5. The maximum and minimum number of visits are given for each route.
6. Total amount of service time for a day must be at most a given upper bound.
7. Those staff capable to handle a particular customer visit may be limited due to their capability and other reasons.
8. Idle time between consecutive visits should be less than a given upper bound.
9. Travel time between two locations depends on the method of transportation.
10. Available transportation modes, either car or bicycle, for each staff are given.
11. Certain customers far away from the station must be visited by car.

3 Set Partitioning Formulation and a Column Generation Heuristic

Set of customer visits is denoted by $N = \{1, 2, ..., n\}$, set of staff, R, and set of assignment of visits to staff $r \in R$, K_r. A constant a_{ki}^r takes value 1 when assignment $k \in K_r$ of staff $r \in R$ includes visit $i \in N$, value 0 otherwise. The total travel time is denoted by c_k^r when the set of customers are visited so that the total travel time is minimized. A variable x_k^r takes value 1 when staff $r \in R$ selects assignment $k \in K_r$.

A home care staff routing and scheduling problem is now formulated:

$$\text{(SPP)} \quad \min \sum_{r \in R} \sum_{k \in K_r} c_k^r x_k^r \tag{1}$$

$$\text{s.t.} \sum_{k \in K_r} x_k^r = 1, \quad r \in R, \tag{2}$$

$$\sum_{r \in R} \sum_{k \in K_r} a_{ki}^r x_k^r = 1, \quad i \in N, \tag{3}$$

$$x_k^r \in \{0, 1\}, \quad r \in R, k \in K_r, \tag{4}$$

Each column of the set partitioning problem corresponds to a feasible route of some staff. The number of possible feasible routes is astronomical, and thus we apply the column generation algorithm to solve the LP relaxation of the set partitioning problem, which provides the lower bound of the optimal objective value. We then solve a set partitioning problem with the generated columns to obtain an upper bound. The column generation subproblem to generate a route of a particular staff with negative reduced cost is a resource-constrained shortest path problem, and we apply the labeling algorithm of Feillet et al. [2].

4 Numerical Experiments

The column generation heuristic was tested on 5 days corresponding to 5 days of a week (Monday through Friday), and their performance was compared with the actual schedule produced by the experienced staff. Performance of the algorithm is evaluated by the value of objective function (total travel time) and by the duality gap. Experiments were performed on a PC with Intel Xeon (2.27 GHz), 12 GB of main memory run on Windows 7 Professional (64 bit). Optimization was performed by AMPL Gurobi Version 4.0.

Instances used and the numerical results are summarized in the upper and lower parts of Table 1, respectively. "Number of soft time windows" in Table 1 means the number of visits whose duration of time window is more than its service time, namely, those "flexible" visits whose start time can be adjusted within the time window. In Table 1, utilization is computed as (total service time + total travel time)/total time, where total time is the time between start and return times of the station.

We make the following observations from the experimental results:

1. Duality gap is roughly less than 1 %, which implies the quality schedules are generated by the algorithm.
2. Computational experiments showed that some instances could be solved very quickly whereas some other instances (such as Monday and Friday) required non-trivial amount of time.
3. Monday and Friday instances take more CPU time, which seems to reflect higher percentage of flexible visits, whose start time can be adjusted within the given time window, for these instances. On the other hand, instances with lower percentage of flexible visits can be solved quickly.
4. Almost 98–99 % of CPU time of these difficult instances is used for the labeling algorithm. The integer program with the generated columns is solved immediately to generate an upper bound.

Table 1 Instances for experiments and numerical results

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of staff	11	10	11	11	12
Number of visits	27	23	22	25	32
Number of soft time windows	17	7	11	7	14
Initial # of columns	6329	1243	3818	4023	5217
# of generated columns	2522	51	59	534	1711
Total # of columns	8851	1294	3877	4557	6928
Total CPU time (s)	5072	22	100	1232	10027
Total travel time (min)	222.8	236.8	226.6	235.4	249.5
Duality GAP (%)	0.66	0.06	0.01	0.2	1.48
Staff utilization (%)	86	90	92	89	91
Actual total travel time (min)	223.4	260.0	264.1	240.0	265.2

5. Naturally total travel time of the generated solution was shorter than that of the the actually used schedule in practice for each day of the week.
6. Utilization of the generated schedule was also higher than that of the actual schedule. This appears to be due to the maximum limit of time between consecutive visits.

5 Algorithm Speed-Up by Pre-processing Input Data

Computational experiments have shown that CPU time to solve some instances must be reduced. A natural approach to reduce CPU time is to improve the algorithm. Considering the fact that some instances are solved quickly, we took an another approach to "make the instance easier" by "reducing the solution space." Reduction of solution space is done by reducing the flexibility inherent in the input data. The idea is based on a natural observation that when a human scheduler faces difficulties in finding a schedule due to too much freedom, he/she often reduces freedom by fixing a part of the schedule one way or other. Fixing is one way to obtaining a feasible schedule, but arbitrary fixing tends to give negative effects on schedule performance. We propose an attempt to pre-process given input data via two simple integer programs (IPs) to reduce freedom of schedules to speed up the algorithm.

Upon checking input data, we decided to focus on (1) a candidate set of staff who could be assigned to each visit (which we call candidate staff), and (2) time window of each visit. In fact, we confirmed that CPU time to solve the problem is reduced substantially by limiting freedom in candidate staff and time window of each visit.

The first simple IP tries to limit candidate staff. The idea behind the model is to make staff i a part of candidate staff for visits of customer j so that the maximum of total service time of individual staff is minimized under the restriction of the

minimum number of candidate staff for visits of customer j, and the minimum number of potential customers assigned to each staff.

Variable x_{ij} is 1 if staff i becomes a part of candidate staff for visits of patient j, 0, otherwise. Those unskilled staff who could handle only a very limited number of customers would be excluded from considerations and we only consider those staff denoted as $R_a \subset R$ who can handle at least moderate number of customers. U denotes the set of customers, a_j^l the smallest number of staff who must be assigned to visits of customer j, b_i^l the smallest number of customers to whom staff i becomes a candidate staff, t_j total service time of visits of customer j, p_i skill level of staff i, q_j minimum skill level required for visits of customer j. Note that skill level of staff i should be at least that of minimum skill level of visit j in order for staff i to be a member of candidate staff for visits of customer j. c_{ij} is 1 if visits of customer j exists during the working hours of staff i, 0 otherwise. Finally, d_j denotes the number of unskilled staff who can handle customer j.

$$\min \quad y \tag{5}$$

$$\text{s.t.} \quad y \geq \sum_{j \in U} x_{ij} t_j, \quad i \in R_a \tag{6}$$

$$\sum_{i \in R_a} x_{ij} + d_j \geq a_j^l, \quad j \in U \tag{7}$$

$$\sum_{j \in U} x_{ij} \geq b_i^l, \quad i \in R_a \tag{8}$$

$$x_{ij} \leq \max(p_i - q_j + 1, 0), \quad i \in R_a, j \in U \tag{9}$$

$$x_{ij} \leq c_{ij}, \quad i \in R_a, j \in U \tag{10}$$

$$x_{ij} \in \{0, 1\}, \quad i \in R_a, j \in U \tag{11}$$

The second simple IP model is developed to narrow down soft time windows. The model tries to reduce the freedom of start time for these flexible visits. The room of flexibility for a visit is the length of original soft time window minus its service time. Reduced time windows of flexible visits would be determined so that room of flexibility of the new time windows is reduced to the user-specified proportion (denoted as β, and called "reduction ratio") of the original room of flexibility. The objective of the model is to minimize the maximum number of visits that may be performed at the same time. Details of the model are omitted due to limited space.

6 Effects of Pre-processing Input Data

To measure the effects of limiting flexibility of solutions using the above two simple IPs, we solve the original scheduling problem (Friday instance) using the original and the pre-processed input data. Both CPU time and the best objective value of the

Table 2 Effects of limiting the flexibility of solutions

$\beta(\%)$	Same as original data	50, 42, 25	45, 42, 25	42, 42, 25	42, 42, 24
100	1.00	1.00	1.00	1.01	1.01
	1.00	0.78	0.26	0.21	0.21
70	1.06	1.06	1.08	1.09	1.09
	1.14	0.31	0.17	0.12	0.12
65	1.05	1.06	1.06	1.06	1.06
	0.31	0.21	0.17	0.08	0.08
60	1.06	1.07	1.09	1.08	1.08
	0.19	0.19	0.11	0.05	0.05
55	1.05	1.11	0.89	–	1.10
	0.37	0.22	0.11	–	0.06

scheduling problem are compared by their ratio. The CPU time ratio of 0.2 indicates that CPU time was 1/5 of the original time after input data pre-processing. Note that both of the two simple IPs for pre-processing can be solved immediately and thus the additional computational burden is negligible and omitted in the ratio calculation. Similarly the objective ratio of 1.05 indicates that the objective value gets worse by 5 % after pre-processing. Note that reducing solution flexibility generally gives adverse effects on the objective value.

Parameters adjusted are the smallest number of customers to whom staff i becomes a candidate staff, i.e., b_i^l in (8), to limit candidate staff, and time window reduction ratio β to narrow down time windows. Parameters b_i^l are set by grouping staff in R_a into 3 groups, (1) administrator who can take care all customers, (2) skilled staff, and (3) other staff. In Table 2, a column headed by 50, 42, 25, e.g., indicates that b_i^l is set to 50 customers for the administrator, 42 and 25 customers for the second and third groups, respectively. Each row corresponds to a particular value of the reduction ratio β. For each combination, the upper entry shows the ratio of objective values of the scheduling model, and the lower entry the ratio of CPU times. For example, for a combination of $b_i^l = (42, 42, 24)$ and $\beta = 0.65$, CPU time is reduced to 8 % of the original CPU time at the cost of 6 % increase in the objective value.

7 Conclusions

A column generation heuristic was developed for home care staff routing and scheduling of a single day. CPU times required to solve the original problems were non-trivial for some days of a week. Pre-processing of input data via two simple IPs was successful to substantially reduce CPU time to solve difficult instances.

References

1. Eveborn, P., Flisberg, P., & Ronnqvist, M. (2006). Laps care—an operational system for staff planning of home care. *European Journal of Operational Research*, *171*, 962–976.
2. Feillet, D., Dejax, P., Gendreau, M., & Dueguen, C. (2004). An exact algorithm for the elementary shortest path problem with resource constraints. *Networks*, *44*, 216–229.
3. Ikegami, A., et al. (2012). Un-you cost wo jyuushi-shita saitekika (In Japanese, Operating-cost-concious optimization). *Communication of Operations Research Society of Japan*, *57*, 695–704.
4. Shao, Y., Bard, J. F., & Jarrah, A. I. (2012). The therapist routing and scheduling problem. *IIE Transactions*, *44*, 868–893.

A Dynamic Customer-Centric Pricing Approach for the Product Line Pricing Problem

Michael Neugebauer

Abstract In this paper, we address a service provider's product line pricing problem for substitutable products. We consider a market that is composed of different customer segments of various sizes. The customers belonging to a segment have the same segment-specific preference rankings. The seller is able to adopt a dynamic pricing strategy and offer different prices for the products to different customer segments. We introduce a mixed-integer linear programming formulation for this problem which is solved by means of IBM ILOG CPLEX. We conduct several computational experiments and present some preliminary results.

1 Introduction and Problem Description

In the literature, the problem of determining optimal prices for a product line has been discussed from multiple points of view and a number of optimization models and procedures have been proposed. Among those, the most prominent models derive from [1–3]. In general, the standard product line pricing problem can be described as follows: A monopolistic seller offers his products $i \in \{1, ..., I\}$ at prices p_i for each product i that are selected from a pre-defined set of price points p_{ia} $(a = 1, \ldots, A_i)$, i.e. $p_i \in \{p_{i1}, \ldots, p_{iA_i}\}$ in order to maximize his revenue. We consider an extended version of this standard problem by addressing a service provider's product line pricing problem for substitutable products in services. The products are sold during a common selling season at the end of which the corresponding services are delivered. During the selling season the seller is allowed to update the prices for the products at points in time $t \in \{0, \ldots, T\}$. The costs of supplying a single unit of a service are not constant but depend on the total amount of service units sold. For this purpose, we

M. Neugebauer (✉)
Department of Analytics and Optimization,
University of Augsburg, Universitätsstr. 16, 86135 Augsburg, Germany
e-mail: michael.neugebauer@wiwi.uni-augsburg.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_44,

Table 1 Products and price points

i	a	p_{ia}
0	**1**	**€0**
1	1	€2
1	**2**	**€5**
2	1	€4
2	**2**	**€6**

Table 2 Preference lists (PL)

$\sigma_s(B)$	PL of segment 1	PL of segment 2
5	Product 1 for €2	Product 2 for €4
4	**Product 1 for €5**	Product 1 for €2
3	Product 2 for €4	**No purchase**
2	**No purchase**	**Product 2 for €6**
1	**Product 2 for €6**	**Product 1 for €5**

introduce piecewise linear cost functions with $l \in \{0, \ldots, L_i\}$ intervals affiliated with the total amount of service units sold. An easy example might be a service provider who has a certain internal capacity at free disposal and is able to buy additional units at a spot market. Finally, we assume that the market is composed of different customer segments with different segment specific preference rankings. The seller can exploit these differences by offering not only time dependent but also segment specific prices for the products. An example might be an online service provider who has identified different customer segments based on their booking history and wants to offer them different prices.

In Sect. 2 we present the demand model. The mathematical modelling is introduced in Sect. 3. After the presentation of some computational experiments in Sect. 4 the paper ends with a conclusion in Sect. 5.

2 Demand Model

Customer behavior is modeled by using a general non-parametric approach (see, e.g., [5]). We consider a market that is composed of different customer segments $s \in \{1, \ldots, S\}$ of various sizes. The customers belonging to a segment have the same preference ranking (or preference list) $\sigma_s(B)$ with $B = \{I \times P_i : i \in \{1, \ldots, I\}\} \cup \{(0, p_0)\} = \{(i, p_i) : i \in \{0, \ldots, I\},\ p_i \in \{p_{i1}, \ldots, p_{iA_i}\}\}$ for all product–price point–combinations (PPPC) including the no-purchase option, "product" $i = 0$ that is always offered at $p_{01} = €0$. $\sigma_s(B) : B \to 1, \ldots, |B|$ is a bijective mapping with the following property: If customer segment s prefers combination $b_1 \in B$ to combination $b_2 \in B$ then $\sigma_s(b_1) > \sigma_s(b_2)$ holds. Based on the decision of the service provider which PPPCs $B' \subseteq B$ to offer, the customers of segment s will choose the combination $b^* = (i^*, p_i^*) \in B'$ with the highest valued $\sigma_s(b^*)$. Preference rankings are assumed not to change during the selling season. In Tables 1 and 2, a short example is given.

There are $I = 2$ products with $A_1 = A_2 = 2$ price points and $S = 2$ customer segments with different preference lists $\sigma_1(B)$ and $\sigma_2(B)$. Assume that the service provider sets the prices $p_{12} = €5$ for $i = 1$ and $p_{22} = €6$ for $i = 2$, i.e. $B' = \{(0, €0), (1, €5), (2, €6)\}$. All available PPPCs B'–including the no-purchase option–are printed in bold face in Tables 1 and 2. In this example, all customers belonging to segment $s = 1$ decide to purchase $i = 1$ for $p_{12} = €5$ because $\sigma_1(1, €5) > \sigma_1(0, €0) > \sigma_1(2, €6)$. All segment 2 customers accordingly choose the no-purchase option $(0, €0)$. We call the set B' a price list. Following this definition, the seller alternatively could have offered one of the following three price lists: $\{(0, €0), (1, €2), (2, €4)\}$, $\{(0, €0), (1, €2), (2, €6)\}$, and $\{(0, €0), (1, €5), (2, €4)\}$. The service provider has the option to define up to K price lists with $K \leq S$. For each customer segment he has to decide which of the price lists $k \in \{1, \ldots, K\}$ he wants to assign to it. The seller is able to adopt a dynamic pricing strategy with a number of price list updates $h \in \{0, \ldots, H\}$ at points in time t with $H < T - 1$. At each price list update the seller has the possibility to reassign price lists to the customer segments. Relating to the example given above, assume that $H = 1$ price list update takes place and $K = 2$ different price lists can be chosen. That means that the service provider can assign different price lists to the two customer segments at the beginning of the selling season ($h = 0$ at $t = 0$). At some point in time during the selling season, the seller is allowed to update these price list assignments.

3 Mathematical Modelling

We introduce a mixed-integer linear programming formulation for this problem. The notation is given below.

Input Parameters:

p_{hkia}:	a^{th} price point of product i $(= p_{ia})$ in price list k at price list update h
σ_{sia}:	preference value of segment s for product i at price point p_{ia}
Δ_{st}:	expected total demand of segment s customers which arrive in $[0, \ldots, t]$
Q_{il}:	end point of interval l of the piecewise linear cost function of service i
m_{il}:	gradient of interval l of the piecewise linear cost function of service i
Δf_{il}:	jump of the piecewise linear cost function of service i at the left endpoint of interval l
M:	sufficiently large number

Decision Variables:

$z_{skh} \in \{0, 1\} = 1$,	if price list k is assigned to segment s at price list update h
$\mu_{ht} \in \{0, 1\} = 1$,	if price list update h takes place in t (with $\mu_{00} = 1$ and $\mu_{H+1,T} = 1$)
$x_{shkia} \in \{0, 1\} = 1$,	if customer segment s chooses product i at price point p_{ia} in price list k after price list update h
$\pi_{hkia} \in \{0, 1\} = 1$,	if product i is offered at price point p_{ia} in price list k after price list update h

$\delta_{il} \in \{0, 1\} = 1$, if the total amount of services i sold extends into interval l
$d_{il} \geq 0$ total amount of services i sold in interval l
$\theta_{shkia} \geq 0$ expected demand of segment s customers which buy product i at price point p_{ia} in price list k and which arrive between price list update h and price list update $h + 1$

The objective function aims to maximize total profits:

$$\text{Max } F(\mathbf{z}, \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\delta}, \mathbf{d}) = \sum_{s=1}^{S} \sum_{h=0}^{H} \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{a=1}^{A_i} \left(\sum_{t=0}^{T} \mu_{h+1,t} \Delta_{st} - \sum_{t=0}^{T} \mu_{ht} \Delta_{st} \right) x_{shkia} p_{hkia} - \sum_{i=1}^{I} \sum_{l=1}^{L_i} (\Delta f_{il} \delta_{il} + d_{il} m_{il}) \; . \tag{1}$$

The revenue is calculated in the first term by multiplying the expected demand of customer segments between two price list updates with the respective price points of the assigned price lists. In the second term, the costs are captured by determining the respective points of the services' cost functions.

To linearize the objective function, we follow the approach by [4] and introduce a continuous auxiliary variable θ_{shkia} such that

$$\theta_{shkia} = \begin{cases} \sum_{t=0}^{T} \mu_{h+1,t} \Delta_{st} - \sum_{t=0}^{T} \mu_{ht} \Delta_{st}, \text{ if } x_{shkia} = 1 \\ 0, \quad \text{otherwise} \end{cases} .$$

The constraints enforcing this linearization are omitted. The remaining constraints can be divided into three groups.

The first couple of constraints, constraints (2) and (3), determine the points in time for the price updates:

$$\sum_{t=0}^{T} \mu_{ht} = 1 \qquad \text{for all } h = 1, \ldots, H \tag{2}$$

$$\sum_{t=0}^{T} \mu_{h+1,t} t \geq \sum_{t=0}^{T} \mu_{ht} (t + 1) \qquad \text{for all } h = 0, \ldots, H - 1 \; . \tag{3}$$

The auxiliary variables μ_{ht} are used to determine the points in time when price list updates take place. Constraints (2) ensure that every price list update is made at exactly one point in time. Note, if less than H price list updates were optimal, the model would determine points in time for the price list updates but the price lists would not change. A chronological ascending order of the price list updates is assured by constraints (3), i.e. price list update $h + 1$ takes place after price list update h.

The second group of constraints, constraints (4–8), are the essential constraints for the product line pricing problem with price points and different price lists.

$$\sum_{k=1}^{K} z_{shk} = 1 \qquad \text{for all } s = 1, \ldots, S,\ h = 0, \ldots, H \tag{4}$$

$$\sum_{i=0}^{I} \sum_{a=1}^{A_i} x_{shkia} = z_{shk} \qquad \text{for all } s = 1, \ldots, S,\ h = 0, \ldots, H,\ k = 1, \ldots, K \tag{5}$$

$$\sum_{a=1}^{A_i} \pi_{hkia} = 1 \qquad \text{for all } h = 0, \ldots, H,\ k = 1, \ldots, K,\ i = 0, \ldots, I \tag{6}$$

$$x_{shkia} \leq \pi_{hkia} \qquad \text{for all } s = 1, \ldots, S,\ h = 0, \ldots, H,\ k = 1, \ldots, K,\ i = 0, \ldots, I,\ a = 1, \ldots, A_i \tag{7}$$

$$\sum_{j=0}^{I} \sum_{p=1}^{A_i} \sigma_{sjp} x_{shkjp} \geq \sigma_{sia} \pi_{hkia} - (1 - z_{shk}) M \qquad \text{for all } s = 1, \ldots, S,\ h = 0, \ldots, H,\ k = 1, \ldots, K,\ i = 0, \ldots, I,\ a = 1, \ldots, A_i\ . \tag{8}$$

The following five sets of constraints hold for all price list updates. One price list is assigned to every customer segment (see constraints (4)). Constraints (5) ensure that each customer segment chooses one PPPC of the price list it is assigned to. Constraints (6) enforce that for every product exactly one price point is chosen in every price list, i.e. all customers get a complete price list for all products. It is ensured by constraints (7) that customers can only choose from PPPCs that are offered. Constraints (8) represent the well-known incentive compatibility constraints: Every customer segment chooses the available PPPC that ranks highest in its preference list. Besides, it is ensured that only the PPPCs of the respective price lists are considered.

The last group of constraints, constraints (9–11), determine the costs of the services that are sold.

$$\sum_{s=1}^{S} \sum_{h=0}^{H} \sum_{k=1}^{K} \sum_{a=1}^{A_i} \theta_{shkia} = \sum_{l=1}^{L_i} d_{il} \qquad \text{for all } i = 1, \ldots, I \tag{9}$$

$$d_{il} \leq \Delta Q_{il} \delta_{il} \qquad \text{for all } i = 1, \ldots, I,\ l = 1, \ldots, L_i \tag{10}$$

$$d_{i,l-1} \geq \Delta Q_{i,l-1} \delta_{il} \qquad \text{for all } i = 1, \ldots, I,\ l = 2, \ldots, L_i\ . \tag{11}$$

Constraints (9) ensure that costs are captured for all services that are sold. In every interval of the linear cost function the total amount of service units sold is restricted by the interval length $\Delta Q_{il} = Q_{il} - Q_{i,l-1}$ with $\Delta Q_{iL_i} = \infty$ for all $i \in \{1, \ldots, I\}$. Furthermore, if the total amount of services extends into an adiacent interval $l + 1$ of the cost function, the total amount of service units sold in the previous interval l is enforced to ΔQ_{il}. That holds for every service (see constraints (10) and (11)).

Table 3 Computation times in seconds

	$K = 1$	$K = 2$	$K = 3$
$H = 0$	< 1	< 1	< 1
$H = 1$	< 1	2	8
$H = 2$	< 1	6	587
$H = 3$	1	12	32
$H = 4$	9	37	> 150000

Table 4 Improvements of total profits

	$K = 1$	$K = 2$	$K = 3$
$H = 0$	–	43.75 %	43.75 %
$H = 1$	59.03 %	70.92 %	70.92 %
$H = 2$	64.84 %	74.13 %	74.13 %
$H = 3$	65.63 %	74.91 %	74.91 %
$H = 4$	65.71 %	74.91 %	74.91 %

4 Computational Experiments

We implemented the mixed-integer linear program in IBM ILOG OPL and solve small instances by means of IBM ILOG CPLEX 12.5 to optimality. All of the tests were performed on a server architecture with an Intel(R) Core(TM) i7 CPU at 2.80 GHz, 8 GB RAM, and Windows 7 Enterprise. In the computational experiments we focus on the effects of some essential contributions to the product line pricing literature on the demand side: We allow the seller to adopt a dynamic pricing strategy and offer different prices for the products to different customer segments. That means, the seller is able to offer K price lists to the customer segments and make H price list updates. We fix the number of customer segments to 5, consider 5 products with 5 price points, a selling season with a length of 12 and piecewise linear cost functions with two intervals. We vary the number of price lists K from 1 to 3 and the number of price updates H from 0 to 4. The computation times are given in Table 3. Referring to the total profits, the model with $K = 1$ and $H = 0$ serves as a benchmark. Table 4 shows the improvements of the total profits in percentage terms.

For our computational experiments, we chose an instance where the flexibility for the seller gained by the incorporation of more price lists and price list updates helps to allocate the demand in a way to avoid the high costs of the spot market. Hence, total profits increase significantly as the number of price lists as well as price list updates is raised. Furthermore, the computing times are increasing significantly as well. As the instances are getting bigger in practice, the service provider has to decide how much (computing) time he wants to invest in order to make his pricing decisions. Furthermore, he must determine how many price lists and price list updates are applicable from a marketing and from a technical point of view.

5 Conclusion

In this paper, we address a service provider's product line pricing problem. Our approach differs from the standard product line pricing problem introduced in the literature in multiple ways. The main contributions on the demand side are the modelling of customer behavior by using preference lists, the incorporation of dynamic pricing and the possibility for the seller to set different prices for the products for different customer segments. On the supply side, the costs of supplying a single unit of a service are not constant but depend on the total amount of service units sold. For this purpose, we introduce piecewise linear cost functions affiliated with each service total amount of units sold.

References

1. Dobson, G., & Kalish, S. (1988). Positioning and pricing a product line. *Marketing Science*, *7*, 107–125.
2. Dobson, G., & Kalish, S. (1993). Heuristics for pricing and positioning a product-line using conjoint and cost data. *Management Science*, *39*, 160–175.
3. Green, P. E., & Krieger, A. M. (1985). Models and heuristics for product line selection. *Marketing Science*, *4*, 1–19.
4. Shioda, R., Tunçel, L., & Myklebust, T. G. J. (2011). Maximum utility product pricing models and algorithms based on reservation price. *Computational Optimization and Applications*, *48*, 157–198.
5. van Ryzin, G., & Vulcano, G. (2011). An expectation-maximization algorithm to estimate a general class of non-parametric choice models. Working paper, Columbia Business School.

Mathematical Formulations for the Acyclic Partitioning Problem

Jenny Nossack and Erwin Pesch

Abstract This paper addresses the problem of partitioning the vertex set of a given directed, edge- and vertex-weighted graph into disjoint subsets (i.e., clusters). Clusters are to be determined such that the sum of the vertex weights within the clusters satisfies an upper bound and the sum of the edge weights within the clusters is maximized. Additionally, the digraph is enforced to partition into a directed, acyclic graph, i.e., a digraph that contains no directed cycle. This problem is known in the literature as acyclic partitioning problem and is proven to be NP-hard in the strong sense. Real-life applications arise, e.g., at rail-rail transshipment yards and in Very Large Scale Integration (VLSI) design. We propose two model formulations for the acyclic partitioning problem, a compact and an augmented set partitioning model.

1 Introduction

Graph partitioning problems are, in general, concerned with the partitioning of the vertex set of an undirected or directed graph into disjoint subsets (also referred to as *clusters*) such that the sum of the edge weights within the clusters is maximized (or equivalently the sum of the edge weights between different clusters is minimized). Most graph partitioning problems are formally defined based on the following framework: Let $G = (V, E)$ denote an undirected (or directed) graph with vertex set $V = \{v_1, \ldots, v_n\}$ and edge set E. We associate with each edge $(v_i, v_j) \in E$ an edge weight $c_{ij} \in \mathbb{R}$ and optionally with each vertex $v_i \in V$ a vertex weight $w_i \in \mathbb{R}$. A *partition* $P = \{V_1, \ldots, V_k\}$ of G is defined as the collection of k disjoint

J. Nossack (✉) · E. Pesch
Department of Management Information Science, University of Siegen, Hölderlinstr. 3, 57068 Siegen, Germany
e-mail: jenny.nossack@uni-siegen.de

E. Pesch
e-mail: erwin.pesch@uni-siegen.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_45,

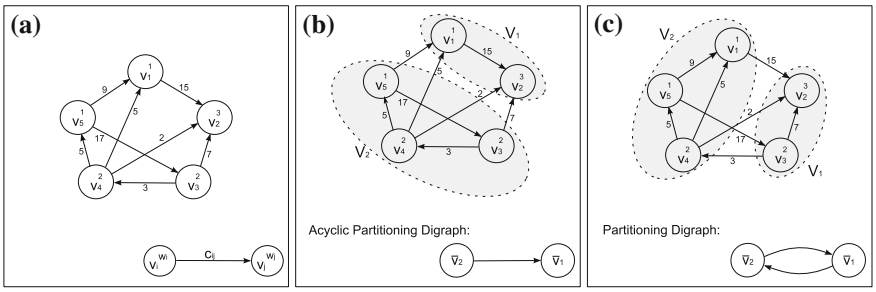


Fig. 1 Example. **a** Digraph. **b** Acyclic partition. **c** Partition

subsets of vertices, $V_1, \ldots, V_k$, such that $\bigcup_{s=1}^{k} V_s = V$ and $V_s \cap V_t = \emptyset$ for all $s, t = 1, \ldots, k$ and $s \neq t$. We refer to V_s, $s = 1, \ldots, k$, as clusters of the partition P. The set of edges connecting vertices of different clusters is called a *cut* and is denoted by $\delta(P) := \{(v_i, v_j) \in E | v_i \in V_s, v_j \in V_t; s, t = 1, \ldots, k; s \neq t\}$. Moreover, the sum of the edge weights defined within clusters and within a cut are denoted as *value*, $val(P) := \sum_{(v_i,v_j)\in E\setminus\delta(P)} c_{ij}$, and *cost*, $cost(P) := \sum_{(v_i,v_j)\in\delta(P)} c_{ij}$, of a partition P, respectively. The graph partitioning problem is to find a partition $P^* = \{V_1, \ldots, V_k\}$ of G such that the value of P^* is maximized (or equivalently the cost of P^* is minimized). Variants of the graph partitioning problem impose side-constraints on this framework.

We consider in this research a constraint variant of the graph partitioning problem, namely the *acyclic partitioning problem*. Given is a directed graph $D = (V, A)$ with vertex weight $w_i \in \mathbb{N}_0^+$ for all $v_i \in V$ and edge weight $c_{ij} \in \mathbb{N}_0^+$ for all $(v_i, v_j) \in A$. We furthermore assume that D is loopless and without multiple edges. Throughout this paper n denotes the number of vertices ($n := |V|$) and m the number of directed edges ($m := |A|$). A partition $P = \{V_1, \ldots, V_k\}$ of digraph D is called an *acyclic partition*, if the sum of the vertex weights of each cluster (also denoted as the *size* of a cluster) is bounded from above by an upper bound $B \in \mathbb{N}_0^+$ and if digraph D partitions into a directed, acyclic graph $D_P = (V_P, A_P)$. We will further refer to D_P as our *partitioning digraph* of the partition P and formally define it as follows: The partitioning digraph D_P includes a vertex for each cluster, i.e., $V_P = (\bar{v}_1, \ldots, \bar{v}_k)$, and defines a directed edge $(\bar{v}_s, \bar{v}_t) \in A_P$ if and only if a directed edge $(v_i, v_j) \in A$ exists in digraph D for any pair of vertices $v_i \in V_s$, $v_j \in V_t$ with $s \neq t$. In summary, the acyclic partitioning problem searches for an acyclic partition $P^* = \{V_1, \ldots, V_n\}$ of D with at most n clusters (i.e., $k := n$) such that the value of P^* is maximized. An example of an acyclic partition and its associated acyclic partitioning digraph is depicted in Fig. 1b for the digraph illustrated in Fig. 1a. Figure 1c illustrates a digraph partition that induces a directed cycle in the partitioning digraph.

The acyclic partitioning problem is—in accordance with most graph partitioning problems—NP-hard in the strong sense [2]. An exact solution algorithm for the acyclic partitioning problem on general digraphs has been presented by [7]. The proposed solution approach is based on a branch-and-bound framework that

integrates constraint propagation. A pseudo-polynomial time solution algorithm for the acyclic partitioning problem on a tree graph topology has been considered by [5]. If all edge weights [4], equivalently, if all vertex weights are equal [2], the acyclic partitioning problem on trees can be solved in polynomial time. Finally, [1] suggest a heuristic procedure for the acyclic partitioning problem on directed graphs with unit vertex and unit edge weights.

2 Properties of the Acyclic Partitioning Problem

In the succeeding section, we will propose two model formulations for the acyclic partitioning problem. These model formulations are based on feasibility conditions formulated in [7]. For the sake of completeness, we will summarize the most relevant feasibility statements and refer the reader to [7] for detailed discussions and proofs.

Theorem 1 [7]

(i) *Let $S_1, \ldots, S_r$ denote the strong components of digraph $D = (V, A)$. If partition $P = \{V_1, \ldots, V_n\}$ is an acyclic partition of D, all vertices that belong to the same strong component lie within the same cluster.*

(ii) *Let $P = \{V_1, \ldots, V_n\}$ denote an acyclic partition of $D = (V, A)$ and let W denote a directed path from $v_i \in V$ to $v_j \in V$. If v_i and v_j belong to cluster V_s, $v_i, v_j \in V_s$, all intermediate vertices on the directed path W belong to V_s.*

(iii) *Let $P = \{V_1, \ldots, V_n\}$ denote an acyclic partition of $D = (V, A)$. If there exists a directed path from $v_i \in V$ to $v_j \in V$ and v_i and v_j belong to different clusters, $v_i \in V_s, v_j \in V_t, s \neq t$, all vertices that can be reached from vertex v_j cannot lie in cluster V_s.*

On account of Theorem 1 (i), the acyclic partitioning problem can be solved on a reduced graph by replacing each strong component by a single vertex and by adjusting the vertex weights. This particular digraph reduction is known in the literature as a *condensation* of a digraph and has the property of being acyclic. We can thus topologically order the vertices of the condensation digraph such that directed edges go from lower-numbered to higher-numbered vertices (refer, e.g., to [7]).

3 Mathematical Formulations

Based on Theorem 1 (i), we assume from now on that digraph D is acyclic with topological ordered vertices. Next, we will present two model formulations for the acyclic partitioning problem, a *compact model* and an *augmented set partitioning model*.

For the compact model, we incorporate four types of decision variables. Binary variable x_{is} denotes whether ($x_{is} = 1$) or not ($x_{is} = 0$) vertex $v_i \in V$ is assigned

to cluster V_s. Furthermore, binary variable z_{ij} indicates if vertices $v_i, v_j \in V, i \neq j$ belong to the same cluster ($z_{ij} = 1$) or not ($z_{ij} = 0$). To ensure that digraph D partitions into an acyclic partitioning digraph $D_P = (V_P, A_P)$, we incorporate a binary decision variable $y_{st} \in \{0, 1\}$ to identify if at least one directed edge $(\bar{v}_s, \bar{v}_t) \in A_P$ is induced between the vertices of clusters V_s and V_t with $s \neq t$ ($y_{st} = 1$) or not ($y_{st} = 0$). The Miller-Tucker-Zemlin subtour elimination constraints, initially introduced for the well-known traveling salesman problem by [6], are applied to the y-variables to enforce the acyclic property of D_P. To formulate the Miller-Tucker-Zemlin constraints, we include an auxiliary variable $\pi_s \in \mathbb{Z}$ for each cluster V_s. Moreover, we introduce parameters p_{ij} which denote whether a directed path exists from vertex v_i to v_j in the digraph D ($p_{ij} = 1$) or not ($p_{ij} = 0$). The model formulation is then given by the following integer programming problem.

$$\max \sum_{(v_i,v_j)\in A} c_{ij} z_{ij} \tag{1}$$

$$\text{s.t.} \sum_{s=1}^{n} x_{is} = 1 \qquad \forall 1 \leq i \leq n \tag{2}$$

$$\sum_{i=1}^{n} w_i x_{is} \leq B \qquad \forall 1 \leq s \leq n \tag{3}$$

$$z_{ij} + x_{is} - x_{js} \leq 1 \qquad \forall 1 \leq i < j \leq n, 1 \leq s \leq n \tag{4}$$

$$x_{is} + x_{jt} - 1 \leq y_{st} \qquad \forall (v_i, v_j) \in A, 1 \leq s \neq t \leq n \tag{5}$$

$$2z_{ij} \leq z_{ih} + z_{hj} \qquad \forall 1 \leq i < h < j \leq n \text{ with } p_{ih} = p_{hj} = 1 \tag{6}$$

$$z_{ih} \leq z_{ij} \qquad \forall 1 \leq i < j < h \leq n \text{ with } p_{ih} = p_{hj} = 1 \tag{7}$$

$$z_{ij} + z_{jh} - z_{ih} \leq 1 \qquad \forall 1 \leq i < j < h \leq n \tag{8}$$

$$z_{ij} - z_{jh} + z_{ih} \leq 1 \qquad \forall 1 \leq i < j < h \leq n \tag{9}$$

$$-z_{ij} + z_{jh} + z_{ih} \leq 1 \qquad \forall 1 \leq i < j < h \leq n \tag{10}$$

$$\pi_s - \pi_t + n y_{st} \leq n - 1 \qquad \forall 1 \leq s \neq t \leq n \tag{11}$$

$$x_{is} \in \{0, 1\} \qquad \forall 1 \leq i \leq n, 1 \leq s \leq n \tag{12}$$

$$z_{ij} \in \{0, 1\} \qquad \forall 1 \leq i < j \leq n \tag{13}$$

$$y_{st} \in \{0, 1\} \qquad \forall 1 \leq s \neq t \leq n \tag{14}$$

$$\pi_s \in \mathbb{Z} \qquad \forall 1 \leq s \leq n \tag{15}$$

Objective function (1) maximizes the value of the digraph partition. Constraints (2) and (3) ensure that each vertex is assigned to exactly one cluster and that the cluster size upper bound B is respected, respectively. Inequalities (4) connect the x-variables and the z-variables. Moreover, constraints (5) connect the x-variables and the y-variables by enforcing y_{st} equal to 1 if at least one directed edge is defined between the vertices of clusters V_s and V_t, and 0 otherwise. Constraints (6) and (7) enforce the cluster to fulfill Theorem 1 (ii) and (iii), respectively. In case of $z_{ij} := 1$, it is enforced by constraints (6) that the intermediate vertices that lie on all directed paths

between vertices v_i and v_j belong to the same cluster. In case of $z_{ij} := 0$, constraints (7) ensure that the vertices which can be reached from vertex v_j cannot lie in the same cluster as vertex v_i. The triangle inequalities (8)–(10), originally proposed by [3] for the clique partitioning problem, verify the transitivity relation: If vertices v_i and v_j belong to the same cluster, as well as vertices v_j and v_h, we may follow that v_i and v_h also belong to this cluster, i.e., if $z_{ij} := 1$ and $z_{jh} := 1$, it follows that $z_{ih} := 1$. The Miller-Tucker-Zemlin constraints are formulated by inequalities (11) and impose the acyclic condition on the partitioning digraph. Finally, constraints (12)–(15) define the domains of the decision variables. Note that the presented formulation is similar to the model proposed in [7]. Instead of using a three-index formulation, we apply a two-index formulation.

For the augmented set partitioning formulation, let Δ denote the set of all clusters. A cluster fulfills Theorem 1 (ii) and (iii), as well as the cluster size capacity B. Let $\sigma^\gamma = (\sigma_i^\gamma | i = 1, \ldots, n)$ denote an incidence vector of a cluster $\gamma \in \Delta$, where $\sigma_i^\gamma = 1$ if vertex $v_i \in V$ is contained in cluster γ and 0 otherwise. The value of a cluster $\gamma \in \Delta$ is denoted by $c_\gamma \in \mathbb{N}_0^+$ and is defined by $c_\gamma := \sum_{(v_i,v_j)\in A} c_{ij}\sigma_i^\gamma \sigma_j^\gamma$. Moreover, three types of decision variables are incorporated in the set partitioning formulation. The binary decision variable θ_γ takes the value 1 if cluster $\gamma \in \Delta$ is part of an acyclic partition and 0 otherwise. To ensure that digraph D partitions into an acyclic partitioning digraph $D_P = (V_P, A_P)$, we incorporate—in accordance to the compact formulation—a binary decision variable $y_{\gamma_s\gamma_t} \in \{0, 1\}$ to identify if at least one directed edge is induced between the vertices of clusters γ_s and γ_t with $\gamma_s, \gamma_t \in \Delta, \gamma_s \cap \gamma_t = \emptyset, s \neq t$. Auxiliary variables $\pi_\gamma \in \mathbb{Z}$ are introduced to formulate the Miller-Tucker-Zemlin subtour elimination constraints. The augmented set partitioning model is then given by the following integer programming model.

$$\max \sum_{\gamma\in\Delta} c_\gamma \theta_\gamma \tag{16}$$

$$\text{s.t.} \sum_{\gamma\in\Delta} \sigma_i^\gamma \theta_\gamma = 1 \qquad \forall 1 \leq i \leq n \tag{17}$$

$$\sigma_i^{\gamma_s}\theta_{\gamma_s} + \sigma_j^{\gamma_t}\theta_{\gamma_t} - 1 \leq y_{\gamma_s\gamma_t} \qquad \forall (v_i, v_j) \in A, \gamma_s, \gamma_t \in \Delta, \gamma_s \cap \gamma_t = \emptyset, s \neq t \tag{18}$$

$$\pi_{\gamma_s} - \pi_{\gamma_t} + n y_{\gamma_s\gamma_t} \leq n - 1 \qquad \forall \gamma_s, \gamma_t \in \Delta, \gamma_s \cap \gamma_t = \emptyset, s \neq t \tag{19}$$

$$\theta_\gamma \in \{0, 1\} \qquad \forall \gamma \in \Delta \tag{20}$$

$$y_{\gamma_s\gamma_t} \in \{0, 1\} \qquad \forall \gamma_s, \gamma_t \in \Delta, \gamma_s \cap \gamma_t = \emptyset, s \neq t \tag{21}$$

$$\pi_\gamma \in \mathbb{Z} \qquad \forall \gamma \in \Delta \tag{22}$$

The objective function (16) corresponds to the maximization of the value of the digraph partition. Constraints (17) ensure that each vertex is assigned to exactly one cluster. Inequalities (18) enforce $y_{\gamma_s\gamma_t}$ equal to 1 if at least one directed edge is defined between the vertices of clusters γ_s and γ_t. The Miller-Tucker-Zemlin constraints are given by inequalities (19). The variable domains are defined by (20)–(22).

Both of these model formulations have their assets and drawbacks. The disadvantages of the compact formulation are twofold. The LP-relaxation of (1)–(15) provides a poor upper bound on the acyclic partitioning problem. A further weakness of the formulation is the variable symmetry in the x-variables. A feasible solution can be represented by $n!$ identical solutions, since the cluster labeling is arbitrary. The poor upper bound, as well as the model symmetry cause LP-based branch-and-bound algorithm, generally implemented in standard solvers like CPLEX, to perform poorly. These disadvantages are discarded by the augmented set partitioning model. However, model (16)–(22) contains a numerous number of rows and columns, such that it is impossible to generate and store the entire constraint matrix for large graph partitioning problems.

4 Conclusion and Future Research

We have presented two model formulations for the acyclic partitioning problem and briefly discussed their characteristics. Future work needs to design algorithms that capture the basic properties of these formulations. The compact model (1)–(15) is to be addressed by solution approaches that reduce the symmetric nature of the suggested model formulation. The augmented set partitioning formulation (16)–(22) points to a solution approach based on column and cut generation. Branch-and-price-and-cut, for instance, is designed for (mixed) integer programming problems in which the constraint matrix contains an enormous number of rows and columns. This method basically combines the strengths of branch-and-price and branch-and-cut. Synthesizing column and row generation is, however, a nontrivial task. A crucial requirement for successful application of this approach is that the structure of the pricing and the separation problem remains unchanged during the algorithm execution.

References

1. Cong, J., Li, Z.,& Bagrodia, R. (1994). Acyclic multi-way partitioning of Boolean networks. In *Proceedings of the 31st Annual Design Automation Conference* (pp. 670–675). New York: ACM.
2. Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability—a guide to the theory of NP-Completeness*. New York: Freeman.
3. Grötschel, M., & Wakabayashi, Y. (1989). A cutting plane algorithm for a clustering problem. *Mathematical Programming*, *45*, 59–96.
4. Hadlock, F. O. (1974). Minimum spanning forests of bounded trees. In *Proceedings of the 5th Southeastern Conference on Combinatorics, Graph Theory, and Computing*, (pp. 449–460). Winnipeg: Utilitas Mathematica Publishing.
5. Lukes, J. A. (1974). Efficient algorithm for the partitioning of trees. *IBM Journal of Research and Development*, *18*, 217–224.

6. Miller, C. E., Tucker, A. W., & Zemlin, R. A. (1960). Integer programming formulation of traveling salesman problems. *Journal of the ACM*, *7*, 326–329.
7. Nossack, J., & Pesch, E. (2014). A branch-and-bound algorithm for the acyclic partitioning problem. *Computers & Operations Research*, *41*, 174–184.

Minimizing Risks for Health at Assembly Lines

Alena Otto

Abstract Reduction of ergonomic risks is one of priorities at assembly lines. In this article, we claim that operational and tactical planning have a significant potential in mitigation of ergonomic risks. We illustrate this on the example of assembly line balancing. For this mid-term planning problem, we show that reduction of ergonomic risks is possible without increasing cycle times or introducing new workstations. We observed that by diversifying tasks assigned to individual workers according to the risks measurement function, we not only balance ergonomic risks among workers, but also achieve their reduction. With help of a two-stage heuristic, developed by us, we were able to find an assembly line balance with *acceptable* risks for *each worker* without increasing the number of stations for about 50 % of instances.

1 Mitigation of Ergonomic Risks Is a Priority Objective at Assembly Lines

It has been done a lot for the well-being of workers since the introduction of assembly lines about two hundred years ago for the manufacturing of muskets [12]. The work has been enriched and workers have been empowered by implementation of team work and quality circles. Various specialized equipment and tools are currently applied to facilitate mounting operations and reduce possibility of failures. However, even nowadays a high prevalence of occupational diseases is observed among assembly line workers. Assembly line workers are especially vulnerable to work-related musculoskeletal disorders, where they take the third place in prevalence after construction workers and health care assistants [8].

Presence of high *ergonomic risks*, or risks for health of workers, is closely connected to the core principles of work organization at assembly lines. The success of

A. Otto (✉)
University of Siegen, Hölderlinstraße 3, Siegen, Germany
e-mail: alena.otto@uni-siegen.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_46,

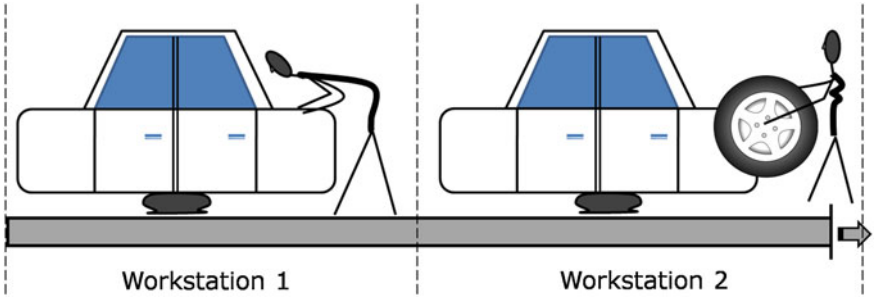


Fig. 1 Illustrative example of an assembly line

assembly lines is based on a high degree of specialization. Thus, each worker receives a *set of tasks* $V = \{1, \ldots, n\}$ to be performed *repetitively* on each workpiece. The workpieces are transported along sequentially arranged workstations manned by one or several workers. As a rule, a workpiece is available at a workstation for a fixed amount of time—cycle time c, and it is moved to the next workstation afterwards. The cycle time is often as low as one minute (e.g., at the final assembly of automobiles) and may even reach 20–40 s. For example, at an assembly line with homogeneous products, cycle time of $c = 20$ s and shift with seven working hours, each worker has to repeat the same set of tasks 1,260 times within a shift.

Therefore, *high repetitivity* of work is the most important physiological *risk factor* at assembly lines. This was also confirmed by a survey conducted among German automotive firms [13]. Because of high repetition frequency, even *moderate* weights and *moderate* levels of force may pose high hazards for health. Indeed, high risks in categories *forces* and *manual material handling* are often detected at assembly line workplaces. Another frequent risk factor is *awkward postures*. For example, mounting operations on undercarriage are performed overhead or above shoulders, whereas tasks in the engine compartment may involve severe bending (see Fig. 1).

Reduction of ergonomic risks is one of high-priority topics in the agenda of both politicians and management of companies. Work-related musculoskeletal disorders bring significant losses to the economy as a whole due to, e.g., lost production output because of the days away of work. Therefore a number of legislative acts and standards were issued to oblige employers to monitor and mitigate ergonomic risks at workplaces (e.g., EU Machinery directive, 2006/42/EC, 89/391/EEC). For factory management, ergonomic risks translate not only into the high compensation costs to be paid to workers, but also into higher failure rate and lower productivity (e.g., [4]). Thus, looking at the failure rates at a Swedish car assembly plant, Eklund [3] found that about 50 % of quality defects stem from (just a few) workplaces with significant ergonomic risks. Because of aging of the workforce in the developed countries, as well as because of an observed tendency for further reduction in cycle times, we expect that ergonomics will receive even more attention in the future.

In [6], we discuss and analyze the ways to reduce ergonomic risks at assembly lines already at the planning stage. The planning stage offers a significant degree of

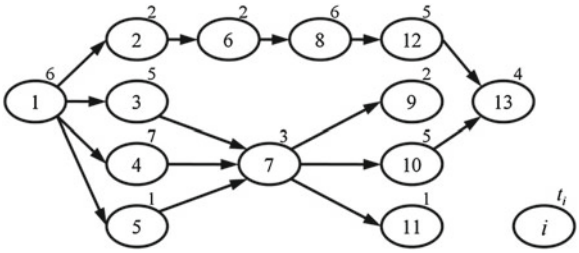


Fig. 2 Illustrative example of a precedence graph

flexibility in designing and changing of processes, therefore ergonomic risks can be mitigated at low costs. Moreover, such improvements often go along with raises in productivity and quality of production.

For this article, we selected one of the important problems of mid-term production planning: assembly line balancing. In the following, we point out the main drivers that can be used to decrease ergonomic risks (Sect. 2). Section 3 describes the model and sketches the solution algorithm. Section 4 reports on major computational results, whereas Sect. 5 provides conclusion and outlook.

2 Drivers Behind the Reduction of Ergonomic Risks

An elegant way to summarize information on tasks at assembly line is the precedence graph $G = (V, E, \mathbf{t})$ (e.g., Fig. 2). Set E summarizes precedence relations between tasks; $(i, j) \in E$ means that task i must be performed before task j. For example, a sit cannot be mounted before the cable has been laid into the undercarriage of the car. Each task $j \in V$ is also characterized by task time t_j, which is calculated according to some of (deterministic) time measurement techniques, established in the industry (e.g., MTM [1]). By assembly line balancing, the industrial engineer partitions a set of tasks V into subsets of tasks S_k, called *station loads*, that have to be performed on each station. Thereby precedence relations must be respected and the sum of task times in each station load should not exceed the cycle time. Below, we refer to such partition as a *feasible balance*. The usual objective is to minimize the number of stations. For example, for precedence graph in Fig. 2 and cycle time $c = 15$, an optimal feasible balance is {{1,2,5,6}, {3,8}, {4,7,12}, {9,10,11,13}} (see Fig. 3).

Ergonomic risks depend on the station load. In many firms, evaluations of ergonomic risks are already performed on routinely basis. A predicate "green" (ergonomic risks are at acceptable level), "yellow" (ergonomic risks are present) or "red" (significant ergonomic risks are present) is assigned to each station depending on this value. Let in our example, station 4 be "red" with task 13 being particularly strenuous. It can be, for example, some operation at undercarriage of a car, which is performed overhead or above shoulder.

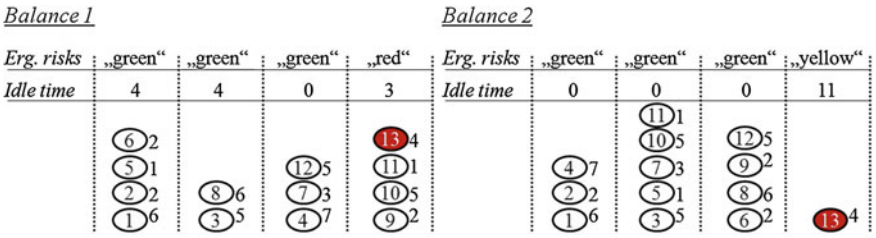


Fig. 3 Example of optimal balances

Even without introducing additional stations, ergonomic risks can be mitigated with help of appropriate assembly line balancing. Three effects can be exploited for this purpose: the targeted distribution of the idle time, the targeted combination of tasks and avoidance of cumulative effects.

The targeted distribution of the idle time. The not scheduled time during the cycle, called idle time $it_k = c - \sum_{j \in S_k} t_j$, provides to the worker opportunity to relax and assume the most favorable posture. The total amount of idle time at the assembly line can be increased only by increasing the number of stations. However, even keeping the number of stations unchanged, we can look for an assembly line balance, where idle time is mostly concentrated at stations containing especially strenuous tasks. In our example, in the first balance, just 20 % of cycle time at "red" station 4 is idle. In balance two, the whole available idle time is concentrated at station 4 to provide to the worker adequate rest time ($it_k = 11$ or 73 % of c). In our example, station 4 turned "yellow".

The targeted combination of tasks. There are several components in evaluation of ergonomic risks. For example, the required energy to perform tasks (physiological component) or whether tasks are perceived as exhausting (psychophysical component). An important role at assembly lines plays biomechanical stress, or forces exerted on musculoskeletal structure. In practice, the risks, which are primarily connected with biomechanical stress, are evaluated independently for each anatomical segment. For example, in OCRA-index [5], which is ergonomic method for estimation of risks for upper extremities, postural index is calculated separately for shoulder, elbow, wrist and hand. Thereby, the worst among these separate indices is taken into the final evaluation of the ergonomic risks. Hence, to provide muscles, tendons and bones appropriate relaxation, we have to avoid combining tasks having exposure on the same anatomical segment, in a single station load.

Avoidance of cumulative effects. In some cases, ergonomic risks increase not linearly in the number of repetitions or in the duration of exposure.

Note, that effect two and effect three help not only to smooth the distribution of ergonomic risks among stations and workers, but also to reduce the total sum of ergonomic risks at the assembly line.

3 Ergonomic Assembly Line Balancing: Model and Solution Procedure

Ergonomic Simple Assembly Line Problem (ErgoSALBP), can be formulated as follows:

$$\text{Min } \Phi = |\{S_k | S_k \neq \emptyset\}| + \omega \cdot \xi(\{S_k\}), \tag{1}$$

$$\text{s.t.} \quad \{S_k\} \text{ is a partition of } V, \text{ which is feasible balance} \tag{2}$$

Ergonomic risks for the whole balance are calculated via an aggregation function ξ. It relies on some ergonomic risk evaluation function to estimate ergonomic risks for a single station, e.g. OCRA-index or EAWS [7]. Further, ξ aggregates ergonomic risks over stations; it may calculate for example, a simple average, some smoothness index or the number of "red" and "yellow" stations.

The objective (1) is to minimize the number of stations and ergonomic risks for the assembly balance as a weighted sum. Depending on the weighting parameter ω, ergonomic risks may be enforced as constraints (for very large values of ω) or reduced as a second-tier objective (for very small values of ω). Note that further industry-specific constraints can be added to the definition of the feasible balance, for example zoning or incompatibility constraints (see [2]).

As a more general version of the simple assembly line balancing problem, ErgoSALBP is NP-hard [10]. Since treating ergonomic objective as a second-tier objective is most relevant for manufacturing [7], we propose a two-stage metaheuristic based on simulation annealing to solve ErgoSALBP.

On the first stage, we find an optimal (or sufficiently near-to-optimum) solution of the correspondent simple assembly line balancing problem, *SALBP* (i.e., we set the expression for Φ in (1) as $|\{S_k | S_k \neq \emptyset\}|$). We fix the number of stations at the found value. Afterwards, we apply simulated annealing metaheuristic to reduce ergonomic risks without introducing additional stations. We generate a neighboring solution by a shift (with probability q) of a task to another station or by swapping (with probability $1 - q$) two tasks. Thereby, we favor selection of tasks from the stations with excessive ergonomic risks. Each time, the incumbent solution is updated, we perform a local search procedure, in order not to miss the (local) optimum. However, we do not change the incumbent solution after the local search procedure is applied; if simulated annealing would be "pushed" into the local optimum, it would be harder for it to overcome the local optimality. The details on the algorithm can be found in [7].

4 Computational Results

As we discussed above, there are several levers that can be used in the planning to reduce ergonomic risks. In our experiments, we investigated whether such improvement is practically meaningful and whether it can be achieved with our algorithm in reasonable time.

For our data set, based on the benchmark data set of Scholl [9] and with ergonomic parameters of tasks similar to those observed and reported in practice, in 90 % of cases, we could achieve an ergonomically better assembly line balance, than the first-stage-solution found by SALOME [11], without increasing the number of stations. For 50 % of instances, a balance with only "green" stations was found.

5 Discussion

Assembly line balancing has a practically meaningful potential to further mitigate ergonomic risks at assembly lines. Currently our approach is being implemented at our cooperation partner.

References

1. Bokranz, R., & Landau, K. (Eds.). (2006). *Produktivitïsmanagement von Arbeitssystemen: MTMHandbuch*. Stuttgart: Schäffer-Poeschel.
2. Boysen, N., Fliedner, M., & Scholl, A. (2007). A classification of assembly line balancing problems. *European Journal of Operational Research*, *183*, 674–693.
3. Eklund, J. A. (1995). Relationships between ergonomics and quality in assembly work. *Applied Ergonomics*, *26*, 15–20.
4. Falck, A. C., Örtengren, R., & Högberg, D. (2010). The impact of poor assembly ergonomics on product quality: A cost-benefit analysis in car manufacturing. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *20*, 24–41.
5. Occhipinti, E. (1989). Ocra: a concise index for the assessment of exposure to repetitive movements of upper limbs. *Ergonomics*, *41*, 1290–1311.
6. Otto, A. (2012). *Application of operational research methods for ergonomic design of working places at assembly lines*. Pro BUSINESS Verlag.
7. Otto, A., & Scholl, A. (2011). Incorporating ergonomic risks into assembly line balancing. *European Journal of Operational Research*, *2012*, 277–286.
8. Schneider, E., & Irastorza, X. (2010). *European risk observatory report. OSH in figures: Work-related musculoskeletal disorders in the EU facts and figures*. Publications Office of the European Union: Luxembourg.
9. Scholl A. (1993) Data of assembly line balancing problems. Schriften zur Quantitativen Betriebswirtschaftslehre 16, TH Darmstadt.
10. Scholl, A. (1999). *Balancing and sequencing of assembly lines*. Heidelberg: Physica.
11. Scholl, A., & Klein, R. (1997). Salome: A bidirectional branch and bound procedure for assembly line balancing. *INFORMS Journal on Computing*, *9*, 319–334.
12. The arsenals of progress (1994). *The Economist*, *330*.
13. Thun, J., Lehr, C. B., & Bierwirth, M. (2011). Feel free to feel comfortable an empirical analysis of ergonomics in the German Automotive Industry. *International Journal of Production Economics*, *133*, 551–561.

A Multi-Objective Online Terrain Coverage Approach

Michael Preuß

Abstract This paper introduces a new multi-objective optimization approach in the field of terrain coverage. With the help of the multi-objective online terrain coverage model, a decentralized autonomous swarm is able to cover an unknown environment. This innovative terrain coverage model has a high impact on autonomous vehicle applications because it considers conflicting objective functions during the coverage process. This important improvement opens up new possibilities for real world applications. The design methodology is based on combining an auction based algorithm with a multiple ant colony optimization route planning algorithm. Experimental analysis is performed on the presented online terrain coverage model which includes the multi-objective route optimization and also a single-objective route optimization. The analysis shows that a multi-objective approach can reduce the repeated coverage and therefore the total coverage time.

1 Introduction

Online terrain coverage models enable an individual agent or a swarm to cover an environment completely without any a priori information about it. A complete coverage requires that every location is visited at least once. This problem is known as *NP*-hard for the multi-agent case [9]. There are different real world applications for [8], search and rescue [3], or military mine hunting [7]. This paper presents a new multi-objective optimization approach in the field of terrain coverage. The studied terrain coverage model can be used for a decentralized autonomous swarm. The self-coordination of the autonomous swarm is organized by an auction based approach [6, 10]. The primary advantage of using an auction based model for an

M. Preuß (✉)
Universität der Bundeswehr München, Fakultät für Informatik,
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
e-mail: michael.preuss@unibw.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_47,

autonomous swarm is the decentralized and robust behavior of the partial subsystems, each agent tries to maximize their individual profit in an opportunistic way. By doing this the global efficiency is increased. This newly introduced multi-objective terrain coverage model enables consideration of conflicting objective functions during the coverage process. This is a significant improvement for real world applications which are faced with multiple objectives like finding the shortest, the safest, the most economical or most informative route. For example there are applications in the fields of logistics, search and rescue or public transportation. In this paper the multiple route planning problem is solved with the help of a multiple ant colony optimization algorithm [1].

This paper is structured as follows. Firstly, the model assumptions and a general overview of the model are presented. Secondly, the use of multiple ant colony optimization for the route planning problem is described. Thereafter the experimental analysis is presented. The last section summarizes the results by using multi-objective optimization in the field of terrain coverage.

2 Online Terrain Coverage Model

Firstly, the model assumptions are presented. The environment is divided into cells, each cell is either free or an obstacle. A free cell can contain a search object. The object will be discovered by visiting the related cell. Each cell is represented by one task. The quadtree decomposition is used for dividing the environment into cells which have the same size as the sensor range [4]. The resulting order of the cells is used to determine the coverage task list. Furthermore there is a search object task list. Both task lists are disjoint.

Every agent is updating both lists individually. If an obstacle is encountered on the agents route, the agent will go left or right around the obstacle in a randomized manor, as long as the agent is on the previous planned route. If the agent circles the obstacle, consequently all tasks represented by the obstacle cells will be deleted. The agents have an obstacle sensor range of one cell in eight directions and in each time unit the agent can move to one of the eight adjacent cells.

Referring to the two task lists, the swarm is separated into two dynamic sub teams. One coverage team which accomplishes coverage tasks and one search object team which discovers object tasks. The maximum size of the search object team has to be defined. For each located object, every unknown adjacent cell will be listed in the object task list. If the agent is a member of the coverage team (search object team), the tasks will be added at the end (to the beginning) of the object task list. The agents can communicate within a defined range and the same mobile ad hoc communication network. Obstacles do not interfere.

Next the self-coordinating auction based approach is explained. As long as an agent does not have an assigned task, auctions will be initiated. If there are unaccomplished object tasks and the search object team is not complete, object tasks will be offered first. Every agent within the same communication network participates

with a bid. Before the agents determine their bids, they exchange information about the environment and accomplished tasks. The agent with the best bid concerning the objective function wins the auction, taking account of the sub teams. Agents which perform a task or have already won other tasks, are including such tasks into the bid.

The determination of the route is implemented in two ways. The implementation of the multi-objective optimization for the multi-objective terrain coverage model (MOOTCM) is explained in Sect. 3. There is a reference single-objective optimization for the multi agent quadtree terrain coverage model (MAQTCM), as well. The aim of the single-objective optimization is the minimization of the route length. Afterwards the quality of each route $G(r_i)$ is evaluated with regard to the sum of the costs and the number of known cells. At this $G(r_i)$ has to be minimized.

3 Multi-Objective Route Planning

The ant colony optimization metaheuristic was developed by Dorigo [2] and which is based on the behavior of an ant colony finding a short route between their lair and a food source. The ants leave behind pheromones on their routes and succeeding ants are attracted. Consequently the probability to follow the same route is increased.

The presented model studies a multi-objective route planning problem. On one hand the route length has to be minimized, whilst on the other hand the maximization of the information gain has to be considered. The information gain is represented by the quantity of unknown cells. For the presented route planning problem the MACO_1 algorithm is used [1]. There are $m + 1$ colonies and m pheromone matrices, where m is the number of objectives. Each single-objective colony discovers the solution space with the help of the referenced pheromone matrix. In addition there is one multi-objective colony which uses all pheromone matrices to find solutions. The pheromone factor $r_S^k(c_{i,j})$ which is considered by the k^{th} single-objective colony encodes information about the objective function f_k. The multi-objective colony considers one of the $r_S^k(c_{i,j})$ pheromone factors randomized. The decision probability for a cell $c_{i,j}$ is determined for the first colony which is trying to find the shortest route by

$$p_S^1(c_{i,j}) = \frac{\left[\frac{r_S^1(c_{i,j})}{\sum_{c_{i,j}\in\Phi} r_S^1(c_{i,j})}\right]^{H_\alpha} \left[\frac{1}{d_S(c_{i,j})}\right]^{H_\gamma}}{\sum_{c_{i,j}^h\in\Phi} \left[\frac{r_S^1(c_{i,j})}{\sum_{c_{i,j}^h\in\Phi} r_S^1(c_{i,j})}\right]^{H_\alpha} \left[\frac{1}{d_S(c_{i,j}^h)}\right]^{H_\gamma}}. \tag{1}$$

The heuristic information d_S is the target range. The neighborhood Φ contains all adjacent cells which have to be considered for the next move. The determined probabilities are used to choose a cell with the Monte-Carlo selection. The second colony which tries to find routes with a high information gain uses a similiar equation in respect to the heuristic information reflecting the priority of a cell, either known or

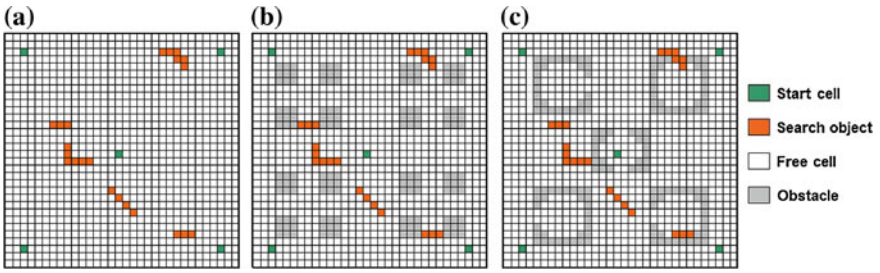


Fig. 1 Environments used for experimentation. In **a** free environment, **b** outdoor-like environment, **c** indoor-like environment

unknown. The pheromone matrices are updated after each cycle. The fitness of the best solution S^k in the current cycle and the overall best solution S^k_{best} of the k^{th} single-objective colony are used to determine the amount of pheromones which are laid on $c_{i,j} \in S^k$. The pheromone matrix r^1 is updated by the first colony with

$$\Delta r^1(c_{i,j}) = \begin{cases} \frac{P_A}{1+Z_\alpha(f_1(S_{best})-f_1(S))}, & \text{if } c_{i,j} \in \mathrm{S}^1 \\ 0, & \text{else} \end{cases} \tag{2}$$

with f_1 denoting the costs of the route. The second colony updates the referring pheromone matrix r^2 dependent on the ratio between the information gain and the costs. The multi-objective colony updates both pheromone matrices r^k with respect to the amount of pheromones for both single-objective colonies.

4 Experiment

To assess the quality of the new approach, experiments were conducted. The experiments run on three well known environments which can be classified as free, outdoor like and indoor like [9]. The initializations used are shown in Fig. 1. Both obstacle environments reduce the total coverage environment by 14 %.

This paper claims to be a proof of concept. Nevertheless, a design of experiments approach [6] was used to determine the following parameter set. The MACO_1 algorithm is used with 120 generations and 80 ants. The best route of the pareto-front is selected with the objective function $G(r_i)$ introduced in Sect. 2. Both weighting factors are 0.5. The heuristic factors are determined for the pheromone value and target range by $H_\alpha = H_\gamma = 0.2$ and the information gain by $H_\beta = 0.6$. The constant factor P_A is set to three. The pheromone evaporation factor is 0.05.

For comparison the single-objective MAQTCM introduced in Sect. 2 and an extended node counting terrain coverage model (NCTCM), are used. The original NCTCM [5] is adapted for a communication structure of the swarm of agents.

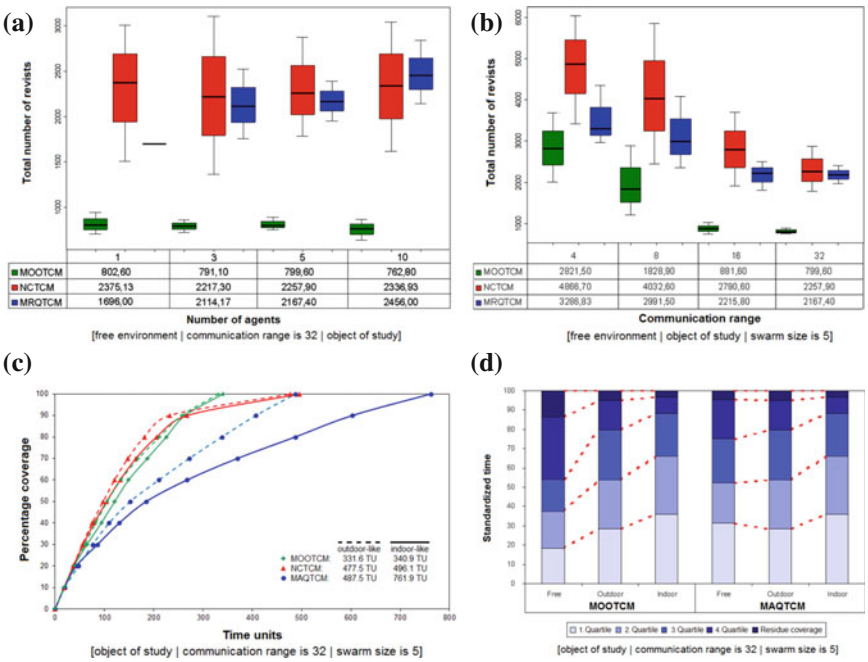


Fig. 2 **a** Influence of the swarm size. **b** Influence of the communication range. **c, d** Influence of the environment

The models are compared with the help of the performance metrics called total coverage time, the percentage of coverage and the total number of revisits. Objects of study are the influences of the obstacle structures, the communication range and the swarm size.

5 Results

The results show the averages of 30 runs for the single-objective MAQTCM and NCTCM. The multi-objective MOOTCM runs only 10 times for each experiment because of the high computational cost.

The swarm size has no significant influence on the number of revisits for the MOOTCM and NCTCM. For the MAQTCM an increasing number of revisits is observable, shown in Fig. 2a. An assumption for the reason is that because of the quadtree decomposition the agents have to cross the center more often.

An increasing communication range has a positive influence on the total coverage time of all three models. More information can be considered for the planning and optimization of the routes. The results show that for the MOOTCM the redundant coverage is increased by 71.7 % for a decreasing communication range by 32 to 4.

In comparison the MAQTCM has an increasing redundant coverage of 34.1 % and the NCTCM of 53.3 %. Nevertheless for each communication range the total coverage time is the best for the MOOTCM. Figure 2b sums up the observations.

The MOOTCM and the NCTCM show no significant differences in the total exploration for the outdoor and indoor like environment. Both models perform slightly better on the outdoor like environment for the first ≈90 % of the coverage process. Thereafter the exploration time converge concerning the indoor like environment. The MOOTCM and NCTCM show a similar exploration behavior. Both are robust towards the analysed obstacle structures. The first results show that the MAQTCM is not robust towards obstacles.

The observation referring to the standardized exploration time of the searching objects shows that there is no significant difference between the MOOTCM and MAQTCM towards the analysed obstacle structures. The MOOTCM and MAQTCM can find the first ≈75 % of the objects faster on the free and outdoor like environment than on the indoor like environment. Furthermore the MOOTCM can find objects fastest on the free environment. Figure 2c, d shows the exploration behaviour for the environment structures.

6 Conclusion

This paper studies a multi-objective terrain coverage approach for a decentralized swarm. For the very first time the multi-objective MOOTCM enables the consideration and adaption of different or even conflicting objective functions during the coverage process. Besides the swarm can exploit the changing environmental conditions in a more flexible way. The analysis shows that the MOOTCM can reduce the repeated coverage and therefore the total coverage time by ≈65 % concerning the node counting NCTCM. For each experiment the MOOTCM yielded the best results. The improvement from the basic single-objective MAQTCM through to the MOOTCM is significant, especially difficult obstacle structures can be treated in a robust way.

Acknowledgments The author would like to thank Silja Meyer-Nieberg and Stefan Pickl for many valuable discussions and comments on this paper.

References

1. Alaya, I., Solnon, C., Ghedira, K. (2007). Ant colony optimization for multi-objective optimization problems. In Proceedings of the 19th ICTAI, 450–457.
2. Dorigo, M. (1992). Optimization, learning and natural algorithms. Ph.D. thesis, Politecnico di Milano, Italy.
3. Kumar, V., Rus, D., & Singh, S. (2004). Robot and sensor networks for first responders. *IEEE Pervasive Computing*, *3*(4), 24–33.

4. Lim, J., & Choo, D. (1998). Sonar based systematic exploration method for an autonomous mobile robot operating in an unknown environment. *Robotica*, *16*(6), 659–667.
5. Pirzadeh, A., & Snyder, W. (1990). A unified solution to coverage and search in explored and unexplored terrains using indirect control. In ICRA (Vol. 3, pp. 2113–2119).
6. Preuß, M. (2011). Terrain coverage: Modelle und Algorithmen. Master's thesis, University of the German Federal Armed Forces, Munich.
7. Sariel, S., Balch, T., & Erdogan, N. (2008). *Naval mine countermeasure missions. IEEE RAM*, *15*(1), 45–52.
8. van Evert, F., van der Heijden, G., Lotz, L., Polder, G., Lamaker, A., de Jong, A., et al. (2006). A mobile field robot with vision-based detection of volunteer potato plants in a corn crop. *Weed Technology*, *20*(4), 853–861.
9. Zheng, X., Koenig, S., Kempe, D., & Jain, S. (2010). Multirobot forest coverage for weighted and unweighted terrain. *IEEE T-RO*, *26*, 1018–1031.
10. Zlot, R., Stentz, A., Dias, M., & Thayer, S. (2002). Multi-robot exploration controlled by a market economy. *IEEE ICRA*, *3*, 3016–3023.

Hot Strip Mill Scheduling Under Consideration of Energy Consumption

Karen Puttkammer, Matthias G. Wichmann and Thomas S. Spengler

Abstract In steel industry hot rolling is an energy-intensive process as steel slabs need to be heated to about 1,250 °C before being rolled on the hot strip mill. Due to time-dependent piecewise energy demand, the total energy consumption for heating is determined by the hot rolling schedule. However, there is no modeling approach known which incorporates the interdependencies between the schedule, the charging time of the slabs, their charging temperature and the energy requirement for heating. We present a MILP formulation for the hot strip mill scheduling problem (HSMSP) under consideration of energy consumption. It takes into account the mentioned interdependencies as well as setups and makespan.

1 Introduction

Hot rolling is one of the most important production processes in steel production. It is the first reshaping process of solid pre-products, so-called slabs. A slab is a steel cuboid that is produced out of liquid steel in the casting process. For reshaping on the hot strip mill the slab needs to be heated in a furnace first. Then the width is reduced on the width compactor and the slab is rolled into a steel strip, that is wound to a coil at the end of the process. Each slab is assigned to a customer order. The customer order determines the final dimensions of the steel strip after hot rolling.

K. Puttkammer (✉) · M. G. Wichmann · T. S. Spengler
Institute of Automotive Management and Industrial Production, Technische Universität Braunschweig, Katharinenstr. 3, 38106 Braunschweig, Germany
e-mail: k.puttkammer@tu-bs.de

M. G. Wichmann
e-mail: ma.wichmann@tu-bs.de

T. S. Spengler
e-mail: t.spengler@tu-bs.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_48,

Hot rolling is an energy-intensive process, which primarily results from the heating process. Because of rising energy prices, the CO_2 certificate trading in the European Union and hard competition due to production overcapacity in the world market, European steel manufacturers strive for a reduction of the specific energy consumption.

The task of production planning at the hot strip mill is to generate a schedule of a given portfolio of production orders. This schedule may consist of one or more rolling turns. Thereby, the specific characteristics of the production process need to be taken into account.

In this paper we present a scheduling model for the introduced planning situation where energy consumption is considered as an objective. Therefore, in Sect. 2 the problem characteristics and the resulting impact on the constraints as well as on the objective function are described in more detail. A modeling approach is discussed in Sect. 3, which incorporates the aspects covered in Sect. 2. An illustrative example is given in Sect. 4. The paper closes with a conclusion and an outlook.

2 Hot Strip Mill Scheduling

In this section the problem characteristics of the hot rolling process are described. In general, there are four characteristics to be considered: slab temperature and energy requirement, the order portfolio, batching in rolling turns with a typical width profile and jumps between adjacent orders.

The first characteristic concerns the slab temperature and the energy requirement. For hot rolling the slabs need to be heated to about 1,250 °C. They are typically charged into the furnaces at a temperature below 500 °C. The temperature results from the previous casting process. The slabs leave the casting process at a temperature of about 1,000 °C. They cool down during the following transportation and storage processes. The cooling curve is a regressive function over time. Thus, the temperature loss is highest in the first hours after casting. The colder a slab is charged into the furnace the more energy is required to reheat it to a given discharging temperature. The charging time of a slab and thereby the energy required for reheating is determined by the production schedule. In the literature there are a few papers that do consider energy consumption in the hot strip mill scheduling problem (HSMSP). Nevertheless, the interdependencies between schedule, charging time, slab age as the time between casting and charging into the furnace, charging temperature and energy requirement are abstracted using highly aggregated cooling assumptions [2, 6, 7]. To cover the interdependencies, the charging time of each slab needs to be calculated dynamically depending on the order sequence.

Second, the order portfolio considered for scheduling contains two kinds of orders. First, there are slabs already waiting in the slab yard for processing. Second, there are virtual slabs. These are slabs en route to the slab yard or slabs which have not been produced yet but have already been scheduled for casting [7]. Either way, virtual slabs must not be scheduled before they are available for hot rolling. The moment a

slab becomes available is determined by the casting time plus the time necessary for transportation to the hot strip mill.

Third, the hot rolling schedule is composed of rolling turns with a characteristic width profile. Due to equipment wear the working rollers of the hot strip mill need to be changed regularly in a setup. The batch of production orders scheduled between two setups is called a turn (also rolling unit or program). The width profile of a turn follows the coffin shape. The coffin shape consists of two sections, the warm-up section and the staple section. In the warm-up section a few slabs are arranged from narrow to wide while in the staple section the slabs are arranged from wide to narrow [4, 5]. In some scheduling approaches the warm-up section is not included into the scheduling task but planned manually [2, 3]. When considering energy consumption, the warm-up section should be part of the planning approach so that the potential of scheduling hot slabs at the beginning of a turn does not get lost. In the warm-up section the number of slabs [7] as well as the width jump between two adjacent orders [4] are limited. Both the cumulated rolling length at the same width within the staple section and the cumulated rolling length of a turn are restricted to ensure surface quality [1, 6].

Fourth, restrictions concerning adjacent orders need to be respected to guarantee good quality with respect to strip flatness. These restrictions include maximal jumps in hardness, in rolling temperature and in thickness where the limits differ for a thickness increase or decrease. Common approaches use penalty cost to allow for these restrictions [1, 7]. For the sake of objectivity of the objective function value they are formulated as hard constraints in our approach.

The objective of scheduling at the hot strip mill is to minimize the production-related cost relevant for the decision. These comprise three types of cost. The first type are setup cost [2, 3]. They involve material cost as the rollers need to be ground after every use. The second type are the cost of energy consumption. Total energy consumption involves the energy required for heating the slabs and efficiency losses. The required energy for one slab equals the specific enthalpy difference of the slab at the discharging and charging temperature multiplied by the slab weight. Given an empiric regressive cooling function and temperature-dependent specific heat capacity, the specific enthalpy difference can be described as a function of the slab age. The resulting enthalpy difference function is concave but can be approximated by piecewise linearization without loss of generality. The third type of cost are opportunity cost. They arise from an extension of the makespan as compared to a minimum necessary time since a longer makespan is related to lost demand. The makespan comprises the time the hot strip mill is occupied due to the scheduled order portfolio. This is the sum of setup times and intermediate charging times. The intermediate charging time of an order is a parameter indicating the time passing between the charging of this order and the succeeding order. Only the time exceeding the minimum makespan is evaluated.

3 Model

In this section the specifics of the mathematical formulation for the scheduling problem are discussed. The objective function of the model is presented and the constraints are described.

The introduced planning problem is an assignment problem. Given an order portfolio of n orders and a schedule with the length of n positions each order i is to be assigned to a position j. Three classes of binary and two classes of continuous variables are used to formulate the model. The binary decision variable x_{ij} indicates whether order i is assigned to position j or not. The binary decision variable y_j is set to one if the warm-up section starts at position j. Equally, z_j indicates the start of the staple section at position j. The continuous decision variable t_i^{charge} describes the charging time, e.g. the processing start of order i. To approximate the specific enthalpy difference function by piecewise linearization, a λ-formulation is used. The continuous decision variable λ_{iu} denotes the nonnegative weight of breakpoint u. A breakpoint is defined by its abscissa a_u and the corresponding function value $f_i(a_u)$. Each specific enthalpy difference value of an order i can then be represented by a linear combination of the breakpoints' function values. Multiplied with the slab weight wg_i it yields the energy demand of the order.

$$MinZ = C_{total} = c_{setup} \cdot \sum_{j=1}^{n} y_j + c_{energy} \cdot \frac{1}{\eta} \cdot \sum_{i=1}^{n} wg_i \cdot \left[\sum_{u=1}^{U} \lambda_{iu} \cdot f_i(a_u) \right] + c_{opp} \cdot \left[\left(\sum_{i=1}^{n} inter_i + \sum_{j=1}^{n} y_j \cdot T^{setup} \right) - LB^{makespan} \right] \quad (1)$$

The objective function (1) minimizes the relevant production-related cost C_{total}, consisting of setup cost, opportunity cost and energy cost. Setup cost arise from the number of setups evaluated with the setup cost factor c_{setup}. Energy cost arise from energy consumption evaluated with the energy cost factor c_{energy}. Energy consumption equals the energy required for heating over all slabs divided by the efficiency factor η. Opportunity cost result from the makespan, calculated by the sum of intermediate times $inter_i$ and setup time T^{setup}. The exceedance of the makespan's lower bound $LB^{makespan}$ is evaluated with the opportunity cost factor c_{opp}.

Eight categories of constraints have to be considered. The first category of constraints defines λ_{iu}. It needs to be ensured that the linear combination of the breakpoints' abscissas equals the age of order i. Further requirements are that at most two adjacent lambdas are greater than zero and that their sum equals one. The second category are assignment constraints. Here, all orders are assigned to a position and vice versa. The third category of constraints defines the charging time. The charging time of an order is obtained by adding the intermediate charging times of all proceeding orders and the incurred setup times. Additionally, virtual slabs must not be scheduled before they are available for rolling. The fourth category of constraints

defines the turn sections within the schedule. They require that the warm-up and the staple section alternate and that at one position either the warm-up or the staple section may start. The fifth category concerns the width profile and the sixth category the jumps between adjacent slabs. They correspond to the requirements specified in Sect. 2. The seventh category of constraints initializes the model. The start of the first warm-up section is fixed at the first position of the schedule. The last category comprises binary and nonnegativity constraints. They define the range of the decision variables.

All the mentioned model constraints can be formulated as linear constraints. Thus, the resulting model can be categorized as a MILP (mixed-integer linear problem). Nevertheless, due to the binary decision variables, the problem is combinatoric and thus hard to solve.

4 Illustrative Example

In this section, the effect of considering energy consumption as an objective is demonstrated. For verification and validation the model was implemented in CPLEX 12.4. A few test instances were solved on a 2.67 GHz CPU with 4GB RAM. Here we present the results of a test instance with 12 orders. The relevant order characteristics are randomly generated based on real world parameter values. The width and thickness, both measured in millimeter, the slab weight measured in tons and the casting time measured in hours with respect to start of production are given in Table 1. The relevant process characteristics are set as follows. The maximal number of orders in the warm-up section is 4 and the maximal increasing (decreasing) width jump is 0.5 mm (0.3 mm). The intermediate time is set very high (1 h) to induce cooling over time. The setup time is set to 1 h, too. All orders are available when planning starts. First, following classical approaches, the setup cost are optimized. Note that because of the deterministic intermediate charging times, the minimization of setups equals the minimization of makespan. Second, the production-related cost relevant for the decision are optimized. In Fig. 1a the solution of the classic approach and in Fig. 1b the solution of the new approach are given. In both solutions two turns are scheduled. Thus, setup cost and opportunity cost are the same in both solutions. However, energy cost are 4.29 % lower when minimizing the production-related cost. The potential for saving energy rises with the length of the planning horizon and the flexibility of assigning orders to alternative positions. Moreover, a small percentage improvement means considerable absolute savings measured against the total energy consumption.

Table 1 Order characteristics of a problem instance with 12 orders

Order	Width	Thickness	Weight	Casting time
1	1,721	2.92	31	−16
2	1,952	1.86	12	−16
3	1,077	1.76	16	−16
4	1,248	2.50	20	−18
5	1,874	2.34	27	−8
6	1,530	1.76	30	−16
7	1,798	2.50	22	−8
8	1,734	2.16	12	−8
9	966	2.00	29	−8
10	1,549	2.50	12	−12
11	1,416	2.50	17	−8
12	1,932	2.26	29	−16

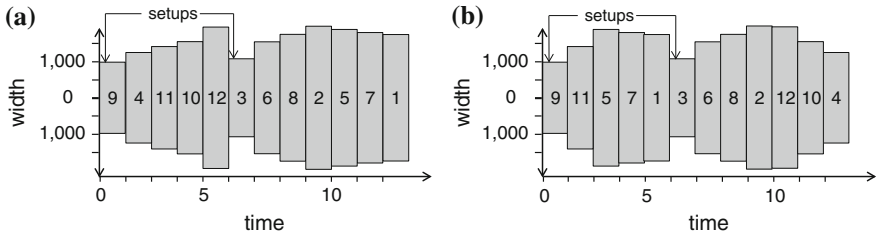


Fig. 1 Optimal solution schedule for problem instance with 12 orders, **a** optimization of setup cost, **b** optimization of total cost

5 Conclusion and Outlook

In this paper we present the HSMSP under consideration of energy consumption. The characteristics of the problem are discussed with focus on the interdependencies between the schedule and the energy requirement for heating. For the first time, a modeling approach that incorporates these characteristics is introduced and a small illustrative example is presented. Further research on adequate solution methods is necessary in order to solve instances of practical problem size within acceptable time.

References

1. Chen, Y. W., Lu, Y. Z., Ge, M., Yang, G. K., & Pan, C. C. (2012). Development of hybrid evolutionary algorithms for production scheduling of hot strip mill. *Computers and Operations Research*, *39*(2), 339–349.

2. Jia, S., Zhu, J., Yang, G., Yi, J., & Du, B. (2012). A decomposition-based hierarchical optimization algorithm for hot rolling batch scheduling problem. *The International Journal of Advanced Manufacturing Technology*, *61*(5), 487–501.
3. Liu, S. (2010). Model and algorithm for hot rolling batch planning in steel plants. *International Journal of Information and Management Sciences*, *21*, 247–263.
4. Tu, N., Luo, X., & Chai, T. (2011). Two-stage method for solving large-scale hot rolling planning problem in steel production. In International Federation of Automatic Control (eds.), *Preprints of the 18th IFAC World Congress* (Vol. 18, pp. 12120–12125).
5. Wang, X., & Tang, L. (2008). Integration of batching and scheduling for hot rolling production in the steel industry. *The International Journal of Advanced Manufacturing Technology*, *36*(5), 431–441.
6. Yadollahpour, M., Bijari, M., Kavosh, S., & Mahnam, M. (2009). Guided local search algorithm for hot strip mill scheduling problem with considering hot charge rolling. *The International Journal of Advanced Manufacturing Technology*, *45*(11), 1215–1231.
7. Zhao, J., Wang, W., Liu, Q., Wang, Z., & Shi, P. (2009). A two-stage scheduling method for hot rolling and its application. *Control Engineering Practice*, *17*(6), 629–641.

Capacitated Network Design

Multi-commodity Flow Formulations, Cutting Planes, and Demand Uncertainty

Christian Raack

Abstract This article provides an overview about the main results and findings developed in the dissertation of the author [8]. In this thesis, we develop methods in mathematical optimization to dimension networks at minimal cost. Given hardware and cost models, the challenge is to provide network topologies and efficient capacity plans that meet the demand for network traffic (data, passengers, freight). We incorporate crucial aspects of practical interest such as the discrete structure of available capacities as well as the uncertainty of demand forecasts. The considered planning problems typically arise in the strategic design of telecommunication or public transport networks and also in logistics. One of the essential aspects studied in this work is the use of cutting planes to enhance solution approaches based on multi-commodity flow formulations. Providing theoretical and computational evidence for the efficacy of inequalities based on network cuts, we extend existing theory and algorithmic work in different directions.

1 Introduction

In this work, we focus on several aspects arising in the context of optimizing and planning the core of nation-wide telecommunication networks. Most of the models and methodology, however, are based on the general notion of *capacitated networks* and *multi-commodity flows* such that the main findings and new approaches are also useful for applications in public transport and logistics. More generally, we were able to enhance some of the most successful approaches to the level of general optimization software handling all kinds of different applications. Solvers such as CPLEX [5], SCIP [10], and GUROBI [4] now scan the problem structure and apply our methods or similar techniques in case they can find network design substructures.

C. Raack (✉)
atesio GmbH, Bundesallee 89, 12161 Berlin, Germany
e-mail: raack@atesio.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_49,

The results in this work have been developed within the German research project EIBONE–*Efficient Integrated Backbone* and the MATHEON project *Integrated planning of multi-layer telecommunication networks* at the *Zuse Institute Berlin*, partially in cooperation with industry partners such as IBM-ILOG, Nokia-Siemens Networks, Deutsche Telekom, and Ericsson.

In the meanwhile, most of the results in this thesis have been published in different Journals: Achterberg and Raack [1], Raack et al. [9], Dash et al. [3], Koster et al. [6] and Poss and Raack [7].

2 The Problem: Capacitated Network Design

The Internet is evolving as the common platform for all classical communication services such as telephony, mailing, and broadcasting TV or radio. Due to its immense flexibility, it has also created new multi-media services, as for instance online-gaming, video-on-demand, (video) instant messaging, and file sharing. This has resulted in an ever increasing demand for higher bit-rates putting pressure on telecommunication network operators to increase network capacity and to efficiently design their infrastructure. In general one has to face the following trade-off: On the one hand, as end-users, we are interested in high *Quality of Service* (QoS), that is, we want fast connections, high throughput, no latency, no packet loss, and no interrupts when using applications that require constant data streams. On the other hand, resources are limited. Network carriers are interested in minimizing capital expenditures (capex) for the necessary technology and equipment but also expenditures for operating the network (opex). In particular, the energy consumption of telecommunication infrastructure has recently moved more into the focus of political and public attention.

This situation creates the classical *capacitated network design problem*: Planning telecommunication networks essentially means to connect locations in a given region and to provide enough capacity at nodes and links in the resulting network in order to meet the demand for bandwidth.

Both in the network engineering community and in the mathematical optimization community, dimensioning networks is known to be extremely challenging already in the setting described so far, that is, the task to create a capacitated network and a network flow supporting a single matrix of traffic demands. However, from the practical point of view, we cannot completely ignore the following additional aspect that gets particular attention in this thesis and even increases the complexity of capacitated network design.

We can never expect to have full knowledge of the traffic demand at the time the design capacity decisions are made. In long-term planning, networks should be dimensioned to meet the future demand. This demand is uncertain. As a consequence, decisions about the actual capacity design are typically made based on traffic estimations, and very often, to avoid bottlenecks and shortages, the traffic is over-estimated.

Over-estimated demand creates over-provisioned networks which in turn results in costly designs and a wastage of resources.

In order to create and operate more resource- and cost-efficient networks the uncertainty of future demand has to be taken into account already in the strategic capacity design process. *Robust network design* tries to address this issue and overcome the mentioned problems. Instead of (over-)estimating a single deterministic traffic scenario, *a set* of realistic traffic scenarios is assumed. Network solutions are then only accepted if they are robust, that is, they are feasible for *all* the considered scenarios.

3 The Methodology: Mixed Integer Programming

To solve different problems in the design of networks we develop techniques in *mixed integer programming*. Demand and capacity constraints are modeled as linear inequalities and equations, while integrality constraints model discrete choices with respect to equipment and/or flow alternatives.

State-of-the-art MIP solvers integrate a cutting plane algorithm into linear programming (LP) based branch-and-bound. There is a trade-off between improving the dual bound by adding more cutting planes and deteriorating the LP re-optimization by adding too many additional constraints which typically slows down the overall algorithm. At this point it is crucial to provide strong cutting planes that cut off large infeasible portions from the relaxation without cutting off feasible points.

Large parts of this thesis concentrate on cutting plane techniques and polyhedral studies applied to network design problems. We provide cutting planes that incorporate the different variables in network design, capacity and flow, and that exploit aspects such as discrete capacity models and demand uncertainty as described above. We thereby study the strength of the developed inequalities theoretically and computationally, that is, we show that the studied inequalities define facets but we also evaluate their algorithmic impact. Our approach is two-fold. We provide cutting plane techniques that can be used to design tailored algorithms to solve specific network design problems. On the other hand, we aim at improving general purpose MIP solvers by including successful special purpose cutting techniques stemming from network design.

Most of the strong inequalities in network design have been derived by studying the problem for very small networks or network substructures. The main idea is to fully understand the mathematical structure and the problem-defining polyhedra for these small instances and to describe (all) important facet-defining inequalities. In a second step, these inequalities are generalized and made available for the original problem.

Following this approach, the inequalities in this thesis are mostly based on *network cuts*. A network cut is a set of links connecting two independent parts of the network, meaning that taking away these links disconnects the network. A *cut-based inequality* essentially states a restriction on the capacity and/or flow on the links defining the

network cut. It might for instance force sufficient cut-capacity. Shrinking each side of a cut to a single node obviously results in a two-node network. In this respect, deriving strong cut-based inequalities is related to understanding network design polyhedra for problems with only two-nodes, also known as *cut-set polyhedra*. We highlight that the concept of studying the facial structure of cut-set polyhedra leads to the well-known and strong cut-set inequalities, flow cut-set inequalities, flow-cover inequalities, or Steiner cut inequalities. We study cut-set polyhedra in different contexts incorporating side-constraints such as demand uncertainty, thereby enhancing and generalizing some of the mentioned cut-based inequalities to these contexts.

4 Main Contributions

This thesis consists of three major parts.

In Part I, we introduce the general concepts and notation. On the one hand we formally introduce the notion of capacity, routing, and multi-commodity flows in networks and describe variations of capacitated network design problems. We present mixed integer programming formulations as well as cutting planes used to tighten the corresponding linear prigramming relaxations. We start with a basic link-flow formulation and integral link capacity variables. We then show how the models can be extended or modified to handle different requirements on the network flow and capacity such as fractional, integral, and single-path flows, unidirectional and bidirectional capacities, as well as multiple link or node capacity modules. For all of these variations we show how to formulate strong cutting planes and review the corresponding literature. The focus is on cut-based inequalities. In this respect, so-called single-node flow sets and cut-set polyhedra are introduced. It is highlighted that cut-based inequalities define facets and can be very effective computationally. That is, the average time to solve network design problems can be reduced substantially and there are many instances that can only be solved in a reasonable amount of time if the mentioned strong inequalities are used as cutting planes.

On the other hand, we focus on solution technology to deal with mixed integer programming formulations in general. As most of the results in this work are related to cutting planes we introduce this methodology in a more general form. We work out how strong inequalities that can be obtained by constraint aggregation techniques in combination with rounding techniques such as mixed integer rounding (MIR). We introduce the concept of complemented mixed integer rounding (c-MIR) as implemented in state-of-the-art MIP solvers. We review crucial known facts but also provide some new insights about the size of MIR aggregations. In particular, we give a short proof of a recent result of Anderson et al. [2] that any MIR cut can be obtained from a subset of linearly independent constraints of the given system.

Part II provides the detailed algorithmic framework of the MCF separator (MCF stands for multi-commodity flow) which combines both areas, that is, successful cutting planes for special purpose network design problems as well as aggregation and separation techniques for general purpose MIP. The MCF separator is now an

integral part of the MIP solvers CPLEX [5] and SCIP [10]. A similar approach based on single-commodity flows and so-called *network inequalities* is now also available in GUROBI (Gu, Chief Technical Officer of GUROBI, 2011, Personal communication). The MCF separator integrates network design specific methodology into these optimization tools which is of particular importance for practitioners that tend to use MIP solvers as black boxes.

The key idea of the MCF-separator is to scan the constraint matrix of general MIP formulations in order to find a substructure that is common to many models for network design problems. This structure consists of a series of similar blocks corresponding to network matrices defining a multi-commodity flow and a coupling of these flow-blocks by capacity constraints. In case of a successful detection, the MCF-separator constructs a network from the obtained information and applies separation methods similar to those introduced in Part I. To obtain inequalities defined on cuts in the detected network, rows of the original system are aggregated accordingly. In this respect, the MCF-separator essentially provides an alternative aggregation framework that is used to provide cut-based base inequalities. These base inequalities are then strengthened by mixed integer rounding. We answer the question of how to detect and construct a network from a multi-commodity flow formulation as well as the question of how to generate valid cut-based inequalities without precisely knowing the network structure. We also report on the computational success of the separator using SCIP and CPLEX. Through extensive computational tests we show that the proposed separation scheme speeds-up the computation for a large set of network design problems by a factor of two on average. Many of these problems can only be solved if the separator is switched on. In roughly 9 % of general MIP instances we find consistent embedded networks and generate violated inequalities. For these instances the computation time is decreased by 18–30 % on average, depending on the solver and test set. For all other instances there is almost no degradation of the optimization performance.

In Part III we study the problem of designing networks without precisely knowing the traffic demand. We discuss how this demand uncertainty can be modeled and review and discuss different demand uncertainty sets. It is shown how the concept of uncertainty affects the methodology to solve capacitated network design problems. Assuming polyhedral uncertainty sets, we highlight that there are different ways of solving the corresponding robust network design problems based on dualization or decomposition techniques. In detailed polyhedral studies we work on the resulting models and robust counterparts. In this respect, we extend the approaches from Part I to the design of robust networks, thereby generalizing and strengthening the strong inequalities from Part I and II. We extend the concept of cut-set polyhedra to robust network design and present facet-defining cut-based inequalities. We provide computational insights comparing different solution approaches, showing progress by separating cutting planes, but also evaluating the robustness of solutions using real-life measurements from IP networks.

Since there is a set of traffic scenarios to be considered, robust network design is a two-stage process. In the first stage we determine capacities. In the second stage we are allowed to change the flow observing realized demands. The flexibility in

the second stage, known as *recourse actions* or *recovery*, can be restricted leading to different routing schemes, static and dynamic routing being the most extensively studied. Following this line, we embed robust network design into the more general framework of two-stage robust optimization with recourse.

The chosen recourse defines a routing scheme which influences the theoretical and computational complexity, but it also influences the price of robustness, that is, depending on the allowed flexibility, the cost for optimal robust network solutions might vary. Static routings are easier to handle computationally as polynomial size reformulations are available. The static routing scheme, however, is very restrictive such that the resulting networks tend to be conservative. Dynamic routings, being the most flexible, produce cheap network designs but lead to hard optimization problems. We introduce a new routing scheme which we call *affine*. Affine routing can be seen as a generalization of static routing allowing for more flexibility. We show that affine routing provides a reasonable alternative in between static and dynamic routing as it still yields polynomial size reformulations. We compare static, affine, and dynamic routing schemes theoretically and discuss their implications. We state necessary and sufficient conditions on polyhedral uncertainty sets under which the three schemes coincide producing the same network cost. Based on realistic network data and demand polytopes, we also compute the cost gap between static, affine, and dynamic solutions. We conclude that for the chosen instances the solutions based on affine routings tend to be as cheap as two-stage solutions with dynamic recourse. In this respect the affine routing principle allows for enough flexibility to almost capture fully flexible dynamic routings. We may hence use affine routing to approximate fully flexible recourse using tractable robust counterparts.

References

1. Achterberg, T., & Raack, C. (2010). The MCF-separator—detecting and exploiting multi-commodity flows in MIPs. *Mathematical Programming C*, *2*, 125–165.
2. Andersen, K., Cornuéjols, G., & Li, Y. (2005). Split closure and intersection cuts. *Mathematical Programming*, *102*(3), 457–493.
3. Dash, S., Günlük, O., & Raack, C. (2011). A note on the MIR closure and basic relaxations of polyhedra. *Operations Research Letters*, *39*(3), 198–199.
4. Gurobi Optimization. GUROBI, 2012
5. IBM. IBM ILOG CPLEX Optimizer, 2012
6. Koster, A. M. C. A., Kutschka, M., & Raack, C. (2011). Robust network design: Formulations, valid inequalities, and computations. *Networks*, *61*(2), 128–149.
7. Poss, M., & Raack, C. (2012). Affine recourse for the robust network design problem: Between static and dynamic routing. *Networks*, *61*(2), 180–198.
8. Raack C. (2012). *Capacitated network design multi-commodity flow formulations, cutting planes, and demand uncertainty*. PhD thesis, TU Berlin.
9. Raack, C., Koster, A. M. C. A., Orlowski, S., & Wessäly, R. (2011). On cut-based inequalities for capacitated network design polyhedra. *Networks*, *57*(2), 141–156.
10. Zuse Institute Berlin. SCIP—Solving Constraint Integer Programs, 2012.

Robustness Analysis of Evolutionary Algorithms to Portfolio Optimization Against Errors in Asset Means

Omar Rifki and Hirotaka Ono

Abstract The Mean-Variance (MV) optimization is a well-studied model for portfolio optimization. Although the main focus is primarily on finding the best efficient portfolios, the MV model is known to be extremely sensitive to perturbations in asset means. This paper investigates the robustness of MV optimization when solved by Evolutionary Algorithms (EA), in the case of linear constraints, i.e., budget constraints and holding constraints. To this end, comparisons were made on Quadratic Programming (QP), Genetic Algorithms (GA) and Evolution Strategies (ES). In order to identify, for EA, robust portfolios, which are supposed to exhibit low sensitivity to small changes in assets means, we proceed by exploiting the population aspect of EA and computing the performance of some selected 'good' individuals under multiple runs subject to perturbations. Comparison of portfolios follows two procedures, the first measures the loss in terms of utility functions, while the second is more practical enabling the decision maker to incorporate a preferred level of robustness. The experimental results using real-world data show that EAs have stronger robustness than QP; many individuals of EA's population outperform the QP-based optimal portfolio.

This work is partially supported by KAKENHI (Nos. 21680001 and 23310104).

O. Rifki (✉) · H. Ono
Department of Economic Engineering, Kyushu University,
Fukuoka 812-8581, Japan
e-mail: omar.rifki.910@s.kyushu-u.ac.jp

H. Ono
e-mail: hirotaka@econ.kyushu-u.ac.jp

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_50,

1 Introduction

The portfolio optimization problem aims to find an optimal allocation of financial capital among a set of available assets. The problem is mostly based on two criteria: minimizing the risk while maximizing the expected return of the investment. Among all the models, the Mean-Variance (MV) analysis has become a classical theoretical framework for portfolio optimization since the pioneering work of Markowitz [7]. Its general formulation according to a risk tolerance level λ for N risky assets can be stated as:

$$\underset{w}{\text{maximize}} \{\mu_p(w) - \lambda\, \sigma_p^2(w)\} \;=\; \sum_{i=1}^{N} w_i \mu_i - \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \sigma_{ij} \tag{1}$$

$$\text{subject to} \sum_{i=1}^{N} w_i \;=\; 1, \quad (C1) \quad \text{and} \quad \forall i \in [|1, N|],\ l_i \leq w_i \leq u_i \quad (C2)$$

where w_i is the weight corresponding to the fraction held in the ith asset. The expected return of the ith asset and the covariance between the returns of the ith and jth assets are respectively denoted by μ_i and σ_{ij}, such that $\sigma_{ii} = \sigma_i^2$ is the variance of the ith asset. Although many hard restrictions can be added to the model such as cardinality constraints, we only consider linear constraints, namely budget constraint $(C1)$, i.e., the entire budget is invested, and holding constraints $(C2)$. In this (linear) case, the MV model is commonly solved by Quadratic Programming (QP) procedures.

In the literature, there is a large number of papers on the benefits and limitations of the MV analysis [8]. One of the salient limitations is that the MV model is quite sensitive to errors in the inputs: means (expected returns), variances and covariances. These factors are de facto containing errors, hence sensitivity is an issue. In fact, as the true future probability distributions of returns are unknown, the used factors are based on statistical estimators, which may contain errors. Best and Grauer [2] explored the sensitivity of optimal portfolios to variations of one asset mean. Their computational results show that a small increase in the mean of one asset can result in large changes in the portfolio's weights, whereas portfolio's return and risk remain slightly affected. On the other hand, Chopra and Ziemba [4] made a distinction between the impact of errors in means, in variances and in covariances. According to the investor's risk tolerance, errors in means can be 10 times or more important than errors in covariances. This value increases with a higher risk tolerance level. Following this result, we focus exclusively on errors of means. Therefore, variances and covariances are supposed exempt from noises.

The sensitive issue does not necessarily imply that the MV framework is flawed. To come across this limitation, researchers considered alterations of the classical model in order to achieve some degree of robustness and reliability mainly by using robust optimization techniques [5]. These techniques aim to incorporate the underlying uncertainty of estimations directly into the optimization process, for instance by

using robust statistical estimators of inputs. The intent of this paper is to consider the sensitivity problem from another angle by using a heuristic solving approach, namely Evolutionary Algorithm (EA), which is a collective term describing a family of stochastic algorithms based on the natural selection principle—*survival of the fittest*. EAs have been remarkably and widely adopted in recent years to solve optimization problems in various domains, in particular those computationally intractable in theoretical sense. In fact, many empirical studies have reported that for MV model, EAs can find good approximate solutions with lower computational costs. The main idea behind these algorithms is to keep evolving a population of candidate solutions one generation after another, using crossover and mutation operators, to hopefully find a global optimum or a suboptimal solution in the worst case.

Outline: The objective of our simulation-based evaluation is to measure accurately how much weight vector w deviates, when assets means are slightly and stochastically perturbed. The proposed robustness measure assess, for QP and two 'dialects' of EA: Genetic Algorithm (GA) and Evolution Strategy (ES), the performance of the *nominal portfolios* under multiple independent noisy runs. We shall refer to nominal portfolios as optimal portfolio allocations computed prior to any perturbation. The first approach measures the loss in terms of utility function, while the second is more pragmatic, allowing the decision maker to incorporate a user-defined robustness tolerance. The matter of robustness assessment will be discussed in detail in the next section. The empirical study and results are presented in Sect. 3, while Sect. 4 concludes this study.

2 Robustness Assessment

Solutions of real-world optimization problems are expected not only to be optimum but also insensitive to small changes affecting the problem variables. Otherwise, a sensitive solution may not be attainable in practice, mainly due to the difficulty of meeting the theoretical assumptions. In this section, we shall present the notion of robustness, and how to assess such concept within the MV framework.

Related work: There are several aspects to evaluate robustness for an optimization process, as indicated in the survey [3]. One important way is to use a robust counterpart of the objective function. This latter one, when used instead of the original function, enhances the robustness of the optimal solution, however, this comes at cost of a performance degradation.[1] This function, so-called *expectation* of the objective function, is sometimes combined with an additional robustness metric also widely used, known as the *variance* of the objective function. Another common method for robustness evaluation is *Sensitivity Analysis* (SA). Once the problem is solved, input parameters are perturbed and output are captured again. If output deviation from the first run is large, this may indicate a lack of robustness.

[1] Robustness and performance are usually conflicting objectives.

Table 1 Nominal portfolios

	QP	EA1	EA2	EA3
(I_0) Nominal case	$EU_{I_0}(w_{QP})$	$EU_{I_0}(w_{EA1})$	$EU_{I_0}(w_{EA2})$	$EU_{I_0}(w_{EA3})$
(I_1) Perturbations 1	$EU_{I_1}(w_{QP})$	$EU_{I_1}(w_{EA1})$	$EU_{I_1}(w_{EA2})$	$EU_{I_1}(w_{EA3})$
…	…	…	…	…
(I_r) Perturbations r	$EU_{I_r}(w_{QP})$	$EU_{I_r}(w_{EA1})$	$EU_{I_r}(w_{EA2})$	$EU_{I_r}(w_{EA3})$
Total				

Adopted Approach: In our simulation-based evaluation, assessments are performed after optimization, as in SA. However, instead of comparing portfolio's weights w directly the output of our model, portfolio's expected utilities $EU(w)$ are examined. We do not refer to *Cash Equivalent* (CE),[2] since all investors of our model have the same quadratic utility function, i.e. $EU(w) = \mu_p(w) - \lambda\ \sigma_p^2(w)$. For our first approach, we start by saving nominal portfolios computed for QP and EA. In case of EA, several nominal individuals w are saved according to a tolerance level α, such that, $EU(w) \geq \alpha\ EU_{\max}$, where $EU_{\max}$ is the best utility achieved in the run, QP and EA included. If negative expected utilities are present, all the $EU(w)$ are scaled by a constant, in order to shift them to the positive domain. Next, we study the behaviour of these saved portfolios under multiple perturbed runs. For each perturbed run, all asset means μ_i are subject to independent standard Gaussian noise according to,

$$\mu_i' = \mu_i + \mu_i * K * X_i, \quad X_i \sim \mathcal{N}(0, 1) \quad \forall i \in [|1, N|],$$

where the parameter K refers to the magnitude of the noises X_i. We have thereafter a new *asset means environment*. We use this environment to recalculate $EU(w)$ for all nominal portfolios, as shown in Table 1. This latter operation is repeated for r different environments of asset means. We shall refer to (I_0) as the nominal run, and (I_1), (I_2), …, (I_r) as the perturbed contexts. The last line of Table 1, counts the number of times each solution, given an environment, is optimal, i.e. best among nominal solutions. Solutions coming up much often might be considered more robust. The second approach, more user-friendly, is shown in Table 2. A solution is accepted or rejected, given a preferred level of robustness β, such that $EU(w) \geq \beta\ EU_{\max}$, e.g. $\beta = 0.95$. The last line of Table 2, counts the number of times each solution is accepted (times of ✓).

[2] CE of a risky portfolio is the certain amount that provides the same utility as the portfolio [4]. Therefore, $U(CE) = EU(w)$. CE is mainly used for comparisons independent of utility units.

Table 2 Preferred level of robustness

	QP	EA1	EA2	EA3
(I_0)	✓	✓	×	×
(I_1)	✓	✓	✓	×
…	…	…	…	…
(I_r)	×	✓	×	×
Total				

3 Experimental Simulations

This paper provides, through real-world data, simulations of a robustness evaluation. Following description of data, methodology and some experimental results.

Data: We consider data from the OR-library [1].[3] It involves stock values of five stock indices, weekly sampled from March 1992 to September 1997. The data used is for the three major stock indices as follows:

Stock index	Country	Number of assets
1. Hang Seng	Hong Kong	31
2. S&P 100	USA	98
3. Nikkei 225	Japan	225

Methods: Picking the appropriate settings for genetic parameters is a hard problem, due to the huge number of possibilities. We have relied on *parameter tuning* for setting these parameters, as described in the tables below. We wrote a Java program using ECJ framework [6] for GA/ES simulations, and ILOG CPLEX 12.5 for QP solving with a smallest convergence tolerance, namely $\varepsilon = 10^{-12}$. Since we noticed that too many nominal EA individuals are stored for the tolerance level $\alpha = 0.99$, only the 10 best EA individuals are picked. For both GA and ES a real-valued representation is adopted. Simulations were run for the risk tolerance $\lambda \in \{2, 3, 4\}$ and for four different values of magnitude $K \in \{0.05, 0.10, 0.15, 0.20\}$. For each λ and K and algorithm, $r = 1{,}000{,}000$ perturbed runs are performed.

GA Parameter	Value	ES Parameter	Value
Population size	300	Population size	300
Generations	1,000	Generations	1,000
Selection	Tournament	Survival selection	$(\mu + \lambda) = (50 + 250)$
Crossover	Simulated binary Crossover (SBX)	Crossover	SBX
Mutation	Polynomial	Mutation	Gaussian
Crossover probability	0.25	Crossover probability	0.25
Mutation probability	0.01	Mutation probability	1 (per gene)

[3] A publicly available benchmark data sets at http://people.brunel.ac.uk/~mastjjb/jeb/info.html.

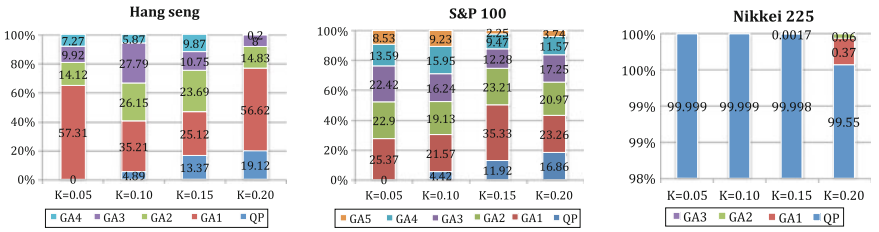


Fig. 1 Repartition of solutions within perturbed runs for GA and $\lambda = 2$

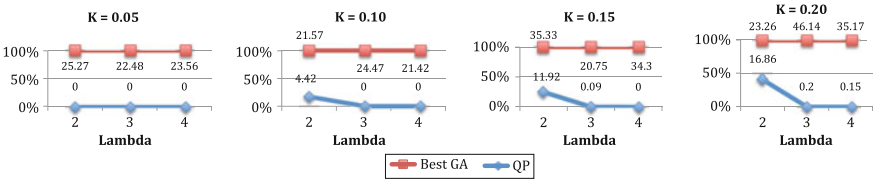


Fig. 2 Varying risk tolerance for GA using S&P index data

Results and discussion: The first set of simulations are based on GA. Figure 1 summarises the results obtained for $\lambda = 2$. These results indicates that for indices with small number of items, Hang Seng and S&P 100, GA-based portfolios provide much better robustness compared with QP solution which is coming up less than 20 % as optimal. The gain in robustness for QP portfolio increases with an increase of the value K, but it is still very low. However, concerning Nikkei 225 index, it is clear that GA-based portfolios are far from being robust. Almost 100 % of the repartitions goes to QP solution. One probable explanation is that GA-based approach does not provide a good solution for the MV problem in case of many assets, as in Nikkei 225. Figure 2 takes into account the variation of the risk tolerance λ. It shows that for the S&P 100 index, QP-based solutions remain too sensitive comparing with GA portfolios, especially for $\lambda \in \{3, 4\}$ where the percentage of QP-based solution being optimal does not exceed 0.2 %. For ES, the following table gives the results for $\lambda = 2$ and $K \in \{0.05, 0.10, 0.15, 0.20\}$, for the three considered indices. These results suggest that ES-based solutions are very sensitive. Only the case of Hang Seng index and $K = 0.20$ exhibits some robustness, with a best ES solution at 15.83 %.

	$K = 0.05$ (%)	$K = 0.10$ (%)	$K = 0.15$ (%)	$K = 0.20$ (%)
Hang Seng	$GP = 99.89$	$GP = 94.65$	$GP = 92.03$	$GP = 70.39$
	$ES1 = 0.1$	$ES1 = 5.3$	$ES1 = 7.81$	$ES1 = 15.83$
		$ES2 = 0.03$	$ES2 = 0.14$	$ES2 = 9.25$
				$ES3 = 1.51$
S&P 100	$GP = 100$ and $ES = 0$			
Nikkei 225	$GP = 100$ and $ES = 0$			

4 Conclusions

In this paper we compared robustness of QP, GA and ES to MV optimization when all assets are slightly and stochastically perturbed. Results using OR-library data [1] show that in case of indices of small assets, many GA individuals outperform the QP-based solution in term of robustness. ES solutions are however more sensitive.

References

1. Beasley, J. E. (1990). OR-Library: Distributing test problems by electronic mail. *Journal of the Operational Research Society*, *41*(11), 1069–1072.
2. Best, M. J., & Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies*, *4*(2), 315–342.
3. Beyer, H. G., & Sendhoff, B. (2007). Robust optimization a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering*, *196*(33), 3190–3218.
4. Chopra, V. K., & Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, *19*(2), 6–11.
5. Fabozzi, F. J., Kolm, P. N., Pachamanova, D., Focardi, S. M., et al. (2007). *Robust portfolio optimization and management*. Wiley, Hoboken.
6. Luke, S., Panait, L., Balan, G., Paus, S., Skolicki, Z., Bassett, J., Hubley, R., & Chircop, A. (2006). *Ecj: A java-based evolutionary computation research system*. Downloadable versions and documentation can be found at the following http://cs.gmu.edu/eclab/projects/ecj
7. Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.
8. Michaud, R. O. (1989). The markowitz optimization enigma: Is optimized optimal? *Financial Analysts Journal*, *45*(1), 31–42.

Two-Stage Robust Combinatorial Optimization with Priced Scenarios

Roman Rischke

Abstract Two-stage robust combinatorial optimization is an established methodology for handling combinatorial optimization problems with uncertain input. Without knowing the actual data, a partial solution needs to be fixed in the first stage which is then extended to a feasible solution in the second stage at higher cost once the data is revealed. The overall goal is to construct a solution that is feasible in all scenarios, i.e., robust against uncertainty, and minimizes the worst-case cost. Since considering all possible scenarios usually leads to a robust solution that is too conservative and too expensive, a central question is to decide on a subset of scenarios to be taken into account. Restricting the set of possible scenarios is a common approach, but this usually depends on subjective decision criteria like the willingness to take risks or the expectation on the future. We propose an alternative concept. Instead of restricting the set of scenarios we price all scenarios, which affects the objective function in such a way that we receive a certain scenario-dependent reward that reduces the overall cost. This leads to new two-stage robust optimization problems. We study complexity and devise approximation algorithms for such problems.

1 Introduction

Practical applications of combinatorial optimization often require decision making under data uncertainty. Reasons for that are usually measurement errors or simply the impossibility of precisely predicting the future. Data uncertainty in optimization problems is usually represented by a set of possible scenarios, where a scenario is a particular realization of the uncertain input parameters. Two-stage robust combinatorial optimization is an established methodology for handling combinatorial optimization problems with uncertain input. The methodology was introduced by

R. Rischke (✉)
Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: roman.rischke@tu-berlin.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_51,

Dhamdhere et al. [3] and subsequently used by different authors [4–7]. Two-stage robust combinatorial optimization is based on a two-stage decision process that we want to illustrate with the following example.

In a frequently flooded region, high restoration cost needs to be paid if a town is under water. In order to avoid these costs, we want to equip some towns with flood protection systems that prevent any damage in case of a flood event. In the first stage we do not know whether a town will be flood-affected or not. That is, the set of flood-affected towns is uncertain in the first stage. We only have information about possible scenarios, where in this example a scenario is a particular set of affected towns. In the second stage a scenario is revealed to us and we have to pay the cost for the restoration of all flood-affected towns that are not equipped with a protection system. We assume, that the cost for the restoration of a town is much higher than the investment in a flood protection system. Our goal is to find a set of towns that we equip with protection systems in the first stage such that we minimize the total cost (first stage cost and second stage cost) in the worst-case scenario. We remark that the worst-case scenario depends on our first stage decision. This means that we want to solve a min-max problem, where we minimize over our possible first stage choices and maximize over our underlying set of scenarios. We can think of a malign adversary, who, once we have taken the first stage decision, picks a scenario which is worst possible with respect to our decision.

We can extend the described two-stage decision process to a decision process with multiple stages. Optimization problems of that kind are called multi-stage robust (combinatorial) optimization problems. However, in this work we restrict our attention to the two-stage model and refer the interested reader to [1, Chap. 14] and the references therein.

2 Priced Scenarios

The goal of robust optimization problems is to construct a solution that is feasible in all scenarios of an underlying scenario set, i.e., robust against uncertainty, and we want to minimize the worst-case cost of the constructed solution. In the above example, feasibility of a solution means that either we have equipped a flood-affected town with a protection system in the first stage or we pay the cost for the restoration of that town in the second stage. If we take all possible scenarios into account, then, by the cost assumption, this will cause us to equip every town with a protection system which is very conservative. Aiming for a more reasonable solution, we firstly have to answer the following central question: *Which scenarios do we take into account?* The more scenarios we take into account the more expensive our robust solution usually is. This is often called “the price of robustness” [2]. Restricting the set of all possible scenarios to a set of reasonable scenarios is a common approach. Note that in robust combinatorial optimization we usually are able to express the set of all possible scenarios in a compact way. In the above example, it is the power set of the set of all considered towns.

In two-stage robust combinatorial optimization, the *discrete scenario approach* (see [3, 8]) and the Γ*-scenario approach* (see [2, 4]) have become the two main approaches for representing the scenario set in the input. In the discrete scenario approach, all scenarios that we want to take into account are explicitly given as part of the input. This approach is appropriate for problems where the number of possible scenarios is manageable. In the Γ-scenario approach, we implicitly describe the scenario set by a parameter Γ which is part of the input. Let us illustrate this by the above example. Suppose, we only want to be robust against situations where at most Γ many towns are flood-affected. One reason for that might be that we expect the number of affected towns to be no greater than Γ, though we do not know the exact set of affected towns. In this case, we only need the set of all towns and the parameter Γ to describe our scenario set. This approach allows us to consider an exponential number of scenarios without listing them all as in the discrete scenario approach.

Both approaches enable us to restrict the set of all possible scenarios, but restricting the scenario set usually depends on subjective decision criteria like the willingness to take risks or the expectation on the future. We propose an alternative approach. Instead of restricting the scenario set we price all scenarios. That is, we define a function that assigns a nonnegative price to each scenario of the unrestricted scenario set. Those prices affect the objective function and lead to new two-stage robust combinatorial optimization problems. We want to illustrate this new approach by the introductory example.

In the new approach, we consider the unrestricted scenario set. In other words, our scenario set is the power set of the set of all considered towns. We extend our example by an insurance company. With this insurance company we negotiate in advance a scenario-dependent price which is paid out to us in the second stage. That is, we agree a price for each scenario that we get paid if the scenario materializes. Our new goal is to find a set of towns that we equip with protection systems in the first stage such that we minimize the balance (first stage cost and second stage cost minus insurance payout) in the worst-case scenario. We assume the insurance premium (fee paid by us to the insurer) to be constant and therefore we can ignore it in the objective function. Again, we want to solve a min-max problem, but now we have a different objective function and we do not have to restrict the scenario set.

Before we give an overview of our results with the new approach, we observe that the new approach generalizes both the discrete scenario approach and the Γ-scenario approach. We can set the price of the scenarios that we want to "exclude" from the scenario set to infinity and thus the adversary has no incentive to choose those scenarios.

3 Results

In this section we give an overview of our results with the new approach. We study complexity and approximation algorithms for a generalization of the afore-mentioned example. This more general problem is called *two-stage robust weighted disjoint*

hitting set problem. The deterministic version of that problem is a special case of many different combinatorial optimization problems, e.g. the *set cover problem* and the *Steiner tree problem*. This also holds for the two-stage robust versions of those problems. Let us first describe the deterministic weighted disjoint hitting set problem (WDHS problem). We are given a set of n elements $E := \{e_1, \dots, e_n\}$, a collection $\mathscr{M} := \{M_1, \dots, M_m\}$ of m pairwise disjoint subsets of E, i.e., $M_i \cap M_j = \emptyset$ for all $i, j \in \{1, \dots, m\}$ with $i \neq j$, and a cost function $c\colon E \to \mathbb{N}$. A feasible solution for the WDHS problem is a set $F \subseteq E$ that has at least one element in common with every set $M \in \mathscr{M}$, i.e., $|F \cap M| \geq 1$ for all $M \in \mathscr{M}$. Thus, the set of feasible solutions can be defined as $\mathscr{F} := \{F \subseteq E \mid \forall M \in \mathscr{M} : |F \cap M| \geq 1\}$. The goal is to find a feasible solution $F \in \mathscr{F}$ that minimizes the total cost $f(c, F) := \sum_{e \in F} c(e)$. The WDHS problem can be solved in polynomial time by selecting the cheapest element out of each set $M \in \mathscr{M}$.

Based on this, we can describe the two-stage robust WDHS problem. As in the deterministic version, we are given the set E, the collection $\mathscr{M}$ and the cost function c as defined above. Additionally, we are given a vector $\lambda := \left(\lambda_{e_1}, \dots, \lambda_{e_n}\right)^{\mathrm{T}} \in \mathbb{Q}^n$ with $\lambda_e \geq 1$ for all $e \in E$ and a scenario set $\mathscr{S}$, where a scenario S is a subset of $\mathscr{M}$. That means that every scenario $S \in \mathscr{S}$ defines a set of feasible solutions $\mathscr{F}^S := \{F \subseteq E \mid \forall M \in S : |F \cap M| \geq 1\}$. In the following, a set $M \in \mathscr{M}$ is called *active* in scenario S if $M \in S$. In the first stage we do not know which scenario $S \in \mathscr{S}$ will materialize in the second stage, but we already can buy elements $e \in E$ in order to "hit" sets. A set $M \in \mathscr{M}$ is *hit* if we buy at least one element of the set M. In the first stage, the cost of an element $e \in E$ is $c(e)$. If we hit a set $M \in \mathscr{M}$ already in the first stage, we do not have to hit M in the second stage in case M is active in the realized scenario. In the second stage a scenario $S \in \mathscr{S}$ is revealed to us and we need to hit all sets $M \in S$ that were not already hit in the first stage. Hitting a set in the second stage is costlier than in the first stage. Every element $e \in E$ has its own given inflation factor $\lambda_e \geq 1$ and costs in the second stage $\lambda_e c(e)$. Let us formulate the goal. We buy a set of elements $F_1 \subseteq E$ already in the first stage and pay $f(c, F_1) := \sum_{e \in F_1} c(e)$. In the second stage we augment the set F_1 by buying an additional set of elements $F_S \subseteq E$, where S is the realized scenario, and we pay $f(\lambda c, F_S) := \sum_{e \in F_S} \lambda_e c(e)$. Our solution (F_1, F_S) is feasible in the scenario S if $F_1 \cup F_S \in \mathscr{F}^S$. The goal is to find a set F_1 and sets F_S, $S \in \mathscr{S}$, such that we minimize the total cost in the worst-case scenario. That is, we want to find a solution for the following min-max problem:

$$\min \left\{ f(c, F_1) + \max_{S \in \mathscr{S}} \{f(\lambda c, F_S)\} \mid \forall S \in \mathscr{S} : F_1 \cup F_S \in \mathscr{F}^S \right\}. \tag{1}$$

Following the discrete scenario approach, we can show that the two-stage robust WDHS problem is *NP*-hard, even if we are given only two scenarios. This is shown by a reduction from the *NP*-complete decision problem *minimum knapsack*. Ideas from the reduction can be used to formulate the problem as a dynamic program that can be solved in pseudo-polynomial time if the cardinality of the scenario set is

constantly bounded from above. By using known methods from two-stage stochastic programming [9], we can show that there is a 2-approximation algorithm for the two-stage robust WDHS problem with discrete scenarios.

However, if we follow the Γ-scenario approach, the two-stage robust WDHS problem can be solved in polynomial time. In this case the scenario set is defined by $\mathscr{S} := \{S \subseteq \mathscr{M} : |S| \leq \Gamma\}$, where $\Gamma \in \{1, \ldots, m\}$ is part of the input. We obtain this result by narrowing down the solution space and greedily selecting sets $M \in \mathscr{M}$ to be hit already in the first stage.

Using the new approach, we slightly need to modify the objective function in (1). We need to incorporate the given price function $p\colon \mathscr{S} \to \mathbb{Q}^+$ that reduces the second stage cost as described in Sect. 2. This leads to the following min-max problem:

$$\min \left\{ f(c, F_1) + \max_{S \in \mathscr{S}} \{f(\lambda c, F_S) - p(S)\} \mid \forall S \in \mathscr{S} : F_1 \cup F_S \in \mathscr{F}^S \right\}, \quad (2)$$

where $\mathscr{S} := 2^{\mathscr{M}}$. In the following, we consider two different price functions p. To define them we need the following additional input. For each set $M \in \mathscr{M}$ we are given a price $\gamma_M \in \mathbb{Q}^+$. Based on this, we say that the price function p is *sum-based* if $p(S) := \sum_{M \in S} \gamma_M$ and we called it *extremum-based* if $p(S) := \max_{M \in S} \gamma_M$ for all $S \in \mathscr{S}$. To explain the following results we introduce the values $\alpha_M := \min_{e \in M} c(e)$ and $\beta_M := \min_{e \in M} \lambda_e c(e)$ for all $M \in \mathscr{M}$.

Let us first consider the case that we are given a sum-based price function p. In this case the two-stage robust WDHS problem (as defined in (2)) can be solved in polynomial time. Let us examine why this is the case and therefor we let $\mathscr{M}' \subseteq \mathscr{M}$ be the collection of all sets that we hit already in the first stage. The collection $\mathscr{M}'$ depends on F_1. First of all, we observe that the adversary will confront us with the worst-case scenario $S^* := \{M \in \mathscr{M} \setminus \mathscr{M}' \mid \beta_M - \gamma_M > 0\}$. Thus we pay $\sum_{M \in \mathscr{M}'} \alpha_M$ in the first stage and $\sum_{M \in S^*} (\beta_M - \gamma_M)$ in the second stage. It is not hard to see that we minimize that payment if we hit all sets $M \in \mathscr{M}$ (at minimum cost) already in the first stage that fulfill $\alpha_M \leq \beta_M - \gamma_M$. This is in line with what intuition tells us.

However, the situation changes drastically if we consider the case that we are given an extremum-based price function p. We can show that the corresponding two-stage robust WDHS problem is *NP*-hard. This is also shown by a reduction from the *NP*-complete decision problem minimum knapsack. We can also show that the problem can be formulated as a dynamic program which can be transformed into an FPTAS by using [10].

4 Conclusion

Pricing scenarios instead of restricting the set of scenarios is an alternative way for dealing with two-stage robust optimization problems. In this work we have motivated and introduced the new approach and we have studied complexity and approximation

algorithms for the two-stage robust WDHS problem. In particular, we have seen that the complexity significantly depends on the pricing method. In this work we have presented our first results with the new approach and we hope to foster further research in this fascinating area.

Acknowledgments I thank my advisors, Rolf H. Möhring and Sebastian Stiller, for all their support and suggestions. I also thank Wiebke Höhn, Daniela Luft and Roland Pörner for their helpful comments on this paper.

References

1. Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization*. *Princeton series in applied mathematics*. Princeton, N.J.: Princeton University Press.
2. Bertsimas, D. J., & Sim, M. (2004). The price of robustness. *Operations Research*, *52*(1), 35–53.
3. Dhamdhere, K., Goyal, V., Ravi, R., & Singh, M. (2005). How to pay, come what may: Approximation algorithms for demand-robust covering problems. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005)* (pp. 367–378).
4. Feige, U., Jain, K., Mahdian, M., & Mirrokni, V. (2007). Robust combinatorial optimization with exponential scenarios. In D.P. Williamson & M. Fischetti (Eds.), *Proceedings of the 12th International Conference on Integer Programming and Combinatorial Optimization (IPCO 2007)*. Lecture notes in computer science, Vol. 4513 (pp. 439–453).
5. Golovin, D., Goyal, V., & Ravi, R. (2006). Pay today for a rainy day: Improved approximation algorithms for demand-robust min-cut and shortest path problems. In B. Durand & W. Thomas (Eds.), *Proceedings of the 23rd Annual Symposium on Theoretical Aspects of Computer Science (STACS 2006)*. Lecture notes in computer science, Vol. 3884 (pp. 206–217).
6. Gupta, A., Nagarajan, V., & Ravi, R. (2010). Thresholded covering algorithms for robust and max-min optimization. In S. Abramsky, C. Gavoille, C. Kirchner, F. Meyer auf der Heide, P.G. Spirakis (Eds.), *Proceedings of the 37th International Colloquium (ICALP 2010)*. Lecture notes in computer science, Vol. 6198 (pp. 262–274).
7. Khandekar, R., Kortsarz, G., Mirrokni, V., & Salavatipour, M. R. (2008). Two-stage robust network design with exponential scenarios. In D. Halperin & K. Mehlhorn (Eds.), *Proceedings of the 16th Annual European Symposium (ESA 2008)*. Lecture notes in computer science, Vol. 5193 (pp. 589–600).
8. Kouvelis, P., & Yu, G. (1997). *Robust discrete optimization and its applications. Nonconvex optimization and its applications*, Vol. 14. Dordrecht: Kluwer Academic Publishers.
9. Shmoys, D. B., & Swamy, C. (2006). An approximation scheme for stochastic linear programming and its application to stochastic integer programs. *Journal of the ACM*, *53*(6), 978–1012.
10. Woeginger, G. J. (2001). *When does a dynamic programming formulation guarantee the existence of an FPTAS?*. Electronic Colloquium on Computational Complexity, Report No. 84.

Workload Balancing in Transportation Crew Rostering

Güvenç Şahin and Fardin Dashty Saridarq

Abstract In crew rostering, balanced workload allocation is a critical issue and an important planning phenomenon that affects both the quality of crew schedules and personnel satisfaction. We focus on workload balancing in transportation systems where deadheading of crew is possible. A network flow formulation of the problem is developed, and an optimal solution method is proposed. We compare the computational performance of the optimal solution method with the solution of the problem with a commercial solver only. We present the results of our computational experiments with well-known problem instances from the crew scheduling literature.

1 Problem Definition

Crew-related costs have a significant share in transportation systems. Especially in railways, this cost constitutes a high portion of the operational expenses. Crew planning at the operational level is concerned with the final assignment of crew members to duties for a finite short planning horizon which is also known as rostering. Rostering is not only concerned with preparing crew schedules that cover all duties but also interested in managerial issues such as fairness in duty assignments and balancing the workload (and associated payments).

Research on workload balancing in transportation crew scheduling is limited. Burke et al. [3] consider the fairness issue as a soft constraint in nurse rostering which ensures distributing duties of various types -morning, night, waiting shifts, etc.- uniformly over the personnel. Bellanti et al. [2] introduce evenly assigned

G. Şahin (✉) · F. Dashty Saridarq
Manufacturing Systems and Industrial Engineering, Sabanci University, Orhanli, Tuzla, 34956 Istanbul, Turkey
e-mail: guvencs@sabanciuniv.edu

F. Dashty Saridarq
e-mail: fardin@sabanciuniv.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_52,

working shifts and days off during the weekends as well as a balanced assignment of morning, afternoon and night shifts as operational requirements. We study the workload balancing problem in transportation crew scheduling, where fairness issue is studied in terms of the workload of crew members.

From a methodological point of view, our work follows the footsteps of both [4, 5]. The balanced bath problem in [4] works with node-disjointness of the path. This idea, however, is not directly applicable on the more generic network representation in [5]. In particular, the existence of deadheading in transportation crew schedules necessiates a further adaptation of the modelling approach in [4] according to the network representation in [5].

2 Mathematical Formulation

The integer programming formulation of the problem is a network flow problem based on the space-time network representation in [5]. In this network, nodes contain time and location information and represent the beginning and ending of events:

- on-duty nodes denote the beginning time and location of a duty;
- tie-up nodes denote the end time and location;
- source node is the origin of all crew members at the home station at the beginning of the planning time horizon;
- a sink node is the final destination of all crew representing the home station at the end of the planning horizon.

Arcs connect the end of events to beginning of other mostly and include six types:

- source arcs emanating from the source node and entering the on-duty nodes at home station represent the origin of crew at the beginning of the planning horizon;
- sink arcs emanating from tie-up nodes and entering the sink node send all crew back to home station at the end of planning horizon;
- duty arcs emanating from an on-duty node and entering a tie-up node represent duties while flow on duty arc represents the coverage of a duty;
- rest arcs represent rest periods which connect a tie-up node to an on-duty node at the same location;
- direct arcs connect two successive duties which have a total time duration less than a predefined time period, these arcs represent the coverage of an excess duty by a crew member where an excess duty covers the first duty, the waiting period between the two duties, and the second duty;
- deadhead arcs from an away tie-up node to a home tie-up node is used to transfer a crew member from the away station to the home station.

A space-time network is accordingly constructed taking into account the rules and restrictions imposed by labor unions, laws and company itself. On this space-time network, a source-sink path is composed of consecutive arcs which represent duties, rest periods and deadheading that correspond to a feasible schedule for a crew

member from the beginning of the planning horizon until the end. As a result, each source-sink path would correspond to a feasible crew schedule. $G = (N, A)$ denotes the network with node set N and arc set A. W_{ij} denotes the workload of arc (i, j); it is zero for all arcs other than duty and deadhead arcs. For a duty arc (i, j), c_{ij} is the number of crew members required to cover the duty. If the sufficient number of crew members to cover all duties in the region is K and the maximum total workload is W_T, then we formulate the corresponding workload balancing problem as follows:

$$\text{Minimize} \quad Z_{\max} - Z_{\min} \tag{1}$$

$$\text{subject to} \quad \sum_{(s,i)\in A} x_{si}^h = 1 \qquad \forall h \in \{1, \ldots, K\} \tag{2}$$

$$\sum_{(i,t)\in A} x_{it}^h = 1 \qquad \forall h \in \{1, \ldots, K\} \tag{3}$$

$$\sum_{(j,i)\in A} x_{ji}^h - \sum_{(i,j)\in A} x_{ij}^h = 0 \qquad \forall h \in \{1, \ldots, K\}, \forall i \in N - \{s, t\} \tag{4}$$

$$\sum_h x_{ij}^h \geq c_{ij} \qquad \forall (i, j) \in A \tag{5}$$

$$\sum_{(i,j)\in A} W_{ij} x_{ij}^h \leq Z_{\max} \qquad \forall h \in \{1, \ldots, K\} \tag{6}$$

$$\sum_{(i,j)\in A} W_{ij} x_{ij}^h \geq Z_{\min} \qquad \forall h \in \{1, \ldots, K\} \tag{7}$$

$$\sum_h \sum_{(i,j)\in A} W_{ij} x_{ij}^h \leq W_T \tag{8}$$

$$x_{ij}^h \geq 0 \qquad \forall h \in \{1, \ldots, K\}, \forall (i, j) \in A \tag{9}$$

where x_{ij}^h denotes the amount of flow over arc (i, j) on a source-sink path h, and $Z_{\max}$ ($Z_{\min}$) denote the maximum (minimum) amount of workload in a schedule corresponding to a source-sink path.

Formulation (1)–(9) finds K source-sink paths corresponding to feasible schedules that are sufficient to cover all duties while minimizing the workload difference ($Z_{\max} - Z_{\min}$) between the maximum workload schedule and the minimum workload schedule.

In a (crew) rostering environment where deadheading is allowed (or indeed necessary), it is possible to add unnecessary workload to crew just for the sake of avoiding the imbalance. However, this increases the total workload, and thus the total costs due to time-based compensation payments. Comparing the network flow problem (1)–(9) with the balanced path problem in [4], it is quite easy to observe the major differences:

- there is no disjointness in our problem as it might limit the addition of necessary deadheads;
- our problem requires that every duty is covered with a given number of sufficient crew members.

In this respect, constraint (8) is crucial as it avoids the addition of unnecessary workload. In the original balanced path formulation in [4], there is no such consideration. W_T can be found by solving a crew scheduling problem as in [5] with the objective of minimizing the total workload with a given number of available crew or heuristically set based on expert opinion.

3 An Exact Algorithm

As even the restricted versions of the problem are NP-hard, we propose a binary search algorithm where each iteration solves a feasibility version of the problem on a smaller feasible region. In essence, we first find a search interval where the optimal value of the objective function (1) lies in. Then, we check if the midpoint value of this interval provides a feasible solution. If so, we bisect this interval and continue with the left half of the interval; otherwise, we continue with the right half of the interval.

In a feasibility check for a given value α, we look for a set of feasible paths (corresponding to a set of feasible schedules) on a sub-graph of the cost expanded version of the network $G = (N, A)$, say $G_q^\alpha = (N_q^\alpha, A_q^\alpha)$ where the shortest source-sink path on G has a cost of q and the longest source-sink path has a cost of $q + \alpha$. If there are K paths covering all duties on G_q^α, then problem is feasible for α. For a fixed binary search interval and α, we may repeat this feasibility check procedure for several values of q starting with smallest possible value.

In Fig. 1, a flowchart for the algorithm is given. We set the initial values of the left hand side (LHS) and right hand side (RHS) values of the search interval to 0 and U, respectively, where U denotes the cost of the longest path on $G = (N, A)$. q_0 denotes the cost of the shortest path on $G = (N, A)$ and F1 refers to the feasibility check procedure. As seen on the flow chart, when F1 does not yield a positive result, the algorithm checks the subgraph with larger cost paths. If F1 does not yield a positive result for a particular α value, then LHS of the search interval is updated. In our implementation of the binary search algorithm, we decrease the computational time of the algorithm by narrowing down the search interval [LHS,RHS]. LHS is replaced by a value $Z_{\max}^{LB} - Z_{\min}^{UB}$ where $Z_{\max}^{LB}$ ($Z_{\min}^{UB}$) denotes a lower (upper) bound for the workload of a feasible schedule with the maximum (minimum) workload schedule. RHS is replaced by a value $Z_{\max}^{UB} - Z_{\min}^{LB}$ where $Z_{\max}^{UB}$ ($Z_{\min}^{LB}$) denotes a lower (upper) bound for the workload of a feasible schedule with the maximum (minimum) workload schedule.

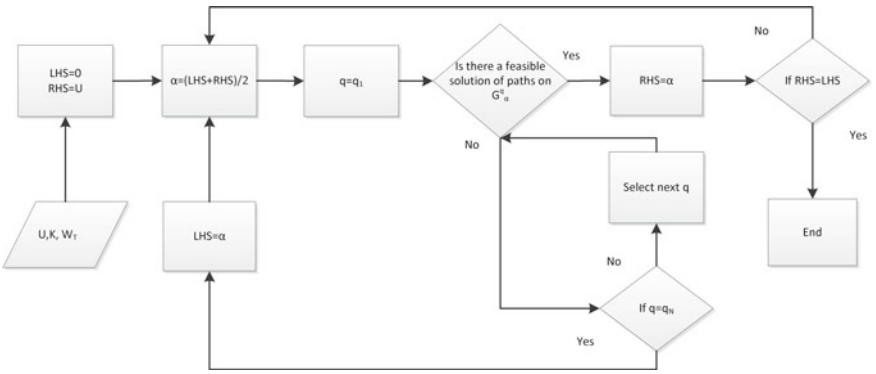


Fig. 1 Flowchart for the binary search algorithm

4 Computational Results and Concluding Remarks

We test the computational performance of the algorithm against the performance of CPLEX in directly solving the integer programming formulation of the problem. Both methods are implemented with C++ while the problems are solved using Concert Technology. Beasley and Cao [1] propose a set of crew scheduling instances which are also used in [4]. In the original instances, any of the duties can be the first or the last duty of a schedule (I). For each instance, we create a limited version (II) where only duties without any predecessor duties can be the first duty of a schedule and only duties without any successor duties can be the last duty of a schedule. Again in the original instances, for any pair of duties, there is a transition cost if it is possible for a crew to perform these two duties consecutively. For each version, we solve one problem with transition costs only and one problem where duties are also attributed with costs proportional to their length in addition to transition costs. As a result, we solve four problem instances generated from an original instance in [1].

Table 1 shows the results. We have set a predetermined time limit of 24 hours for the solution of a problem. In the problem name, the first field denotes the number of duties, the second field denotes the number of paths (K), the third field denotes the version of the problem (I or II) and the last field shows if duties also have costs (c). Under the CPLEX heading, we show the minimum workload ($Z_{\min}$) and maximum workload ($Z_{\max}$) along with the difference (OFV) in the final solution by CPLEX and the computation time (Time). Under the Binary Search Algorithm heading, we show the parameter values that specify the search interval and the final solution (OFV) along with the computational time (Time).The exact optimal solution by CPLEX fails in two instances. In '80-20-II-c', CPLEX terminates with insufficient memory. In '100-20-I-c', the time limit is reached with no feasible solution found. The binary search algorithm narrows down the interval to [485,492] in '80-20-II-c' while it fails to find any solution in '100-20-II-c'. In essence, neither of these methods dominate the other.

Table 1 Results

Problem	CPLEX				Binary Search Algorithm					
	Z_{min}	Z_{max}	OFV	Time	Z_{min}^{LB}	Z_{min}^{UB}	Z_{max}^{LB}	Z_{max}^{UB}	OFV	Time
50-13-I	0	993	993	9 s	0	0	993	993	993	14 s
50-13-II	307	993	686	14 s	307	307	993	1,190	686	11 s
50-13-I-c	161	1,660	1,499	24 s	161	161	1,660	1,705	1,499	1.11 m
50-13-II-c	712	1,660	948	28 s	572	712	1,660	1,794	948	2.36 h
80-20-I	0	536	536	3.16 m	0	0	536	647	536	3.56 m
80-20-II	229	737	508	20.56 m	202	229	737	737	508	27.31 m
80-20-I-c	34	1,194	1,160	42.6 m	34	34	1,194	1,299	1,160	2.25 m
80-20-II-c	–	–	–	–	512	734	1,222	1,401	[485,492]	L:24 h
100-20-I	0	658	658	6.93 m	0	0	658	804	658	2.01 h
100-20-II	321	990	669	39.15 m	229	321	990	990	669	2.81 h
100-20-I-c	–	–	–	L:24 h	34	34	1,574	1,672	1,540	33.45 m
100-20-II-c	1,030	1,596	566	5.38 h	707	1,030	1,596	1,805	–	L:24 h

We focus on the workload balancing problem in transportation systems where deadheading is used. In the previous work by [4], the problem where there is no option for deadheading is investigated as an example for balanced path problems. Our mathematical formulation demonstrates that the problem is significantly different when deadheading is considered. In addition, our early computational results on the same problem instances used in [4] show that the problem is highly difficult from a computational point of view. As some instances cannot even be solved in reasonable time, it requires development of specialized algorithms or heuristic methods.

Acknowledgments This research has been supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant 110M495.

References

1. Beasley, J. E., & Cao, B. (1996). A tree search algorithm for the crew scheduling problem. *European Journal of Operational Research*, *94*, 517–526.
2. Bellanti, F., Carello, G., & Tadei, R. (2004). A greedy-based neighborhood search approach to a nurse rostering problem. *European Journal of Operational Research*, *153*, 28–40.
3. Burke, E., Cowling, P., Causmaecker, P. D., & Berghe, G. V. (2001). A memetic approach to the nurse rostering problem. *Applied Intelligence*, *15*, 199–214.
4. Cappanera, P., & Scutellá, M. (2011). Color-coding algorithms to the balanced path problem: computational issues. *INFORMS Journal on Computing*, *23*, 446–459.
5. Şahin, G., & Yüceoğlu, B. (2011). Tactical crew planning in railways. *Transportation Research Part E: Logistics and Transportation Review*, *47*, 1221–1243.

An Optimal Placement of a Liaison with Short Communication Lengths Between Two Members of the Same Level in an Organization Structure of a Complete K-ary Tree

Kiyoshi Sawada

Abstract This paper proposes a model of placing a liaison which forms relations to two members in the same level of a pyramid organization structure when lengths between the liaison and the other members are less than those between members except the liaison in the organization such that the communication of information between every member in the organization becomes the most efficient. For a model of adding a node of liaison which gets adjacent to two nodes with the same depth in a complete K-ary tree of height H where the lengths of edges between the liaison and the other members are $L(0 < L < 1)$ while those of edges between members except the liaison are 1, an optimal pair of two nodes to which the node of liaison gets adjacent is obtained by maximizing the total shortening distance which is the sum of shortening lengths of shortest paths between every pair of all nodes in the complete K-ary tree.

1 Introduction

The pyramid organization structure can be expressed as a rooted tree, if we let nodes and edges in the rooted tree correspond to members and relations between members in the organization respectively. Then the pyramid organization structure is characterized by the number of subordinates of each member, that is, the number of children of each node and the number of levels in the organization, that is, the height of the rooted tree [3, 7]. Moreover, the path between a pair of nodes in the rooted tree is equivalent to the route of communication of information between a pair of members in the organization, and adding edges to the rooted tree is equivalent

K. Sawada (✉)
Department of Policy Studies, University of Marketing and Distribution Sciences,
3-1 Gakuen-nishi-machi, Nishi-ku, Kobe 651-2188, Japan
e-mail: Kiyoshi_Sawada@red.umds.ac.jp

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_53,

to forming additional relations other than that between each superior and his direct subordinates [6].

Liaisons [2] which have roles of coordinating different sections are also placed as a means to become effective in communication of information in an organization. We have proposed some models of placing a liaison which forms relations to members in the same level of a pyramid organization structure which is a complete K-ary ($K = 2, 3, \ldots$) tree of height $H (H = 2, 3, \ldots)$ [4, 5]. When a node of liaison which gets adjacent to nodes with the same depth is placed, an optimal depth is obtained by minimizing the sum of lengths of shortest paths between every pair of all nodes in the complete K-ary tree. These models are expressed as all edges have the same length. However, we should consider that edges between the liaison and the other members are shorter than those between members except the liaison in the organization.

This paper proposes a model of placing a liaison which forms relations to two members in the same level of a pyramid organization structure which is a complete K-ary tree of height H when lengths between the liaison and the other members are less than those between members except the liaison in the organization. The lengths of edges between the liaison and the other members are $L (0 < L < 1)$ while those of edges between members except the liaison are 1. This paper obtains an optimal pair of two members to which the liaison forms relations such that the communication of information between every member in the organization becomes the most efficient. This means to obtain an optimal pair of two nodes to which the node of liaison gets adjacent minimizing the sum of lengths of shortest paths between every pair of all nodes when an added node of liaison gets adjacent to two nodes with the same depth of a complete K-ary tree of height $H (H = 1, 2, \ldots)$. A complete K-ary tree is a rooted tree in which all leaves have the same depth and all internal nodes have K children [1].

If $l_{i,j} (= l_{j,i})$ denotes the distance, which is length of the shortest path from a node v_i to a node v_j in the complete K-ary tree of height H, then $\sum_{i<j} l_{i,j}$ is the total distance. Furthermore, if $l'_{i,j}$ denotes the distance from v_i to v_j after getting adjacent in the above model, $l_{i,j} - l'_{i,j}$ is called the shortening distance between v_i and v_j, and $\sum_{i<j} (l_{i,j} - l'_{i,j})$ is called the total shortening distance. Minimizing the total distance is equivalent to maximizing the total shortening distance.

2 Formulation of Total Shortening Distance

This section formulates the total shortening distance when a node of liaison is added and gets adjacent to two nodes with the same depth $N (N = 1, 2, \ldots, H)$ in a complete K-ary ($K = 2, 3, \ldots$) tree of height $H (H = 1, 2, \ldots)$. The lengths of edges between the node of liaison and the two nodes to which the node of liaison gets adjacent are $L (0 < L < 1)$ while those of edges between nodes except the node of liaison are 1. Since we don't consider efficiency of communication of information

between the liaison and the other members, the total shortening distance doesn't include the shortening distance between the node of liaison and the other nodes in a complete K-ary tree.

The node of liaison can get adjacent to two nodes with the same depth N of a complete K-ary tree in N ways that lead to non-isomorphic graphs. Let $R_H(N, D)$ denote the total shortening distance by getting adjacent to two nodes, where $D(D = 0, 1, 2, \ldots, N-1)$ is the depth of the deepest common ancestor of the two nodes to which the node of liaison gets adjacent. For the case of $D = 0$, the total shortening distance is denoted by $S_H(N)$. Since getting adjacent to two nodes shortens distances only between pairs of descendants of the deepest common ancestor of the two nodes to which the node of liaison gets adjacent, we obtain

$$R_H(N, D) = S_{H-D}(N - D). \tag{1}$$

We formulate $S_H(N)$ in the following. Let v_0^X and v_0^Y denote the two nodes to which the node of liaison gets adjacent and assume that $D = 0$. Let v_k^X and v_k^Y denote ancestors of v_0^X and v_0^Y, respectively, with depth $N-k$ for $k = 1, 2, \ldots, N-1$. The sets of descendants of v_0^X and v_0^Y are denoted by V_0^X and V_0^Y respectively. (Note that every node is a descendant of itself [1].) Let V_k^X denote the set obtained by removing the descendants of v_{k-1}^X from the set of descendants of v_k^X and let V_k^Y denote the set obtained by removing the descendants of v_{k-1}^Y from the set of descendants of v_k^Y, where $k = 1, 2, \ldots, N-1$.

Since getting adjacent to two nodes doesn't shorten distances between pairs of nodes other than between pairs of nodes in $V_k^X (k = 0, 1, 2, \ldots, N-1)$ and nodes in $V_k^Y (k = 0, 1, 2, \ldots, N-1)$, the total shortening distance can be formulated by adding up the following three sums of shortening distances:

1. The sum of shortening distances between every pair of nodes in V_0^X and nodes in V_0^Y.
2. The sum of shortening distances between every pair of nodes in V_0^X and nodes in $V_k^Y (k = 1, 2, \ldots, N-1)$ and between every pair of nodes in V_0^Y and nodes in $V_k^X (k = 1, 2, \ldots, N-1)$.
3. The sum of shortening distances between every pair of nodes in $V_k^X (k = 1, 2, \ldots, N-1)$ and nodes in $V_k^Y (k = 1, 2, \ldots, N-1)$.

The sum of shortening distances between every pair of nodes in V_0^X and nodes in V_0^Y is given by

$$A_H(N) = 2\{M(H-N)\}^2 (N - L), \tag{2}$$

where $M(h)$ denotes the number of nodes of a complete K-ary tree of height $h(h = 0, 1, 2, \ldots)$. The sum of shortening distances between every pair of nodes in V_0^X and nodes in $V_k^Y (k = 1, 2, \ldots, N-1)$ and between every pair of nodes in V_0^Y and nodes in $V_k^X (k = 1, 2, \ldots, N-1)$ is given by

$$B_H(N) = 4M(H-N)\sum_{i=1}^{N-1}\{(K-1)M(H-i-1)+1\}(i-L), \tag{3}$$

and the sum of shortening distances between every pair of nodes in $V_k^X (k = 1, 2, \ldots, N-1)$ and nodes in $V_k^Y (k = 1, 2, \ldots, N-1)$ is given by

$$\begin{aligned} C_H(N) = 2\sum_{i=1}^{N-2}&\{(K-1)M(H-i-2)+1\} \\ &\times \sum_{j=1}^{i}\{(K-1)M(H-N+j-1)+1\}(i-j-L+1), \end{aligned} \tag{4}$$

where we define $\sum_{i=1}^{0} \cdot = 0$ and $\sum_{i=1}^{-1} \cdot = 0$. From the above equations, the total shortening distance $S_H(N)$ is given by

$$S_H(N) = A_H(N) + B_H(N) + C_H(N). \tag{5}$$

3 An Optimal Depth D^* for Each Depth N

This section shows an optimal depth D^* of the deepest common ancestor of the two nodes which maximizes $R_H(N, D)$ for each depth N. From Eqs. (1) and (5) we have

$$\begin{aligned} R_H(N,D) = {}& 2\{M(H-N)\}^2(N-D-L) \\ &+ 4M(H-N)\sum_{i=1}^{N-D-1}\{(K-1)M(H-D-i-1)+1\}(i-L) \\ &+ 2\sum_{i=1}^{N-D-2}\{(K-1)M(H-D-i-2)+1\} \\ &\times \sum_{j=1}^{i}\{(K-1)M(H-N+j-1)+1\}(i-j-L+1). \end{aligned} \tag{6}$$

Theorem 1 *$D^* = 0$ maximizes $R_H(N, D)$ for each N.*

Proof If $N = 1$, then $D^* = 0$ trivially. If $N \geq 2$, then $D^* = 0$ since

$$\begin{aligned} &R_H(N, D+1) - R_H(N, D) \\ &\quad = -2\{M(H-N)\}^2 - 4M(H-N)\{(K-1)M(H-N)+1\}(N-D-L-1) \end{aligned}$$

$$
\begin{aligned}
&- 4M(H-N)\sum_{i=1}^{N-D-2}(K-1)\{M(H-D-i-1)-M(H-D-i-2)\}(i-L)\\
&- 2\sum_{i=1}^{N-D-3}(K-1)\{M(H-D-i-2)-M(H-D-i-3)\}\\
&\times\sum_{j=1}^{i}\{(K-1)M(H-N+j-1)+1\}(i-j-L+1)\\
&- 2\{(K-1)M(H-N)+1\}\sum_{j=1}^{N-D-2}\{(K-1)M(H-N+j-1)+1\}\\
&\times(N-D-L-j-1)\\
&< 0 \qquad (7)
\end{aligned}
$$

for $D = 0, 1, 2, \ldots, N-2$. □

Theorem 1 shows that the most efficient way of forming relations to two members in each level is that to two members which doesn't have common superiors except the top.

Since the number of nodes of a complete K-ary tree of height h is $M(h) = \left(K^{h+1}-1\right)/(K-1)$, $S_H(N)$ of Eq. (5) becomes

$$
\begin{aligned}
S_H(N) = \frac{1}{(K-1)^3}\Big\{&(-2NL+2N)K^{2H-N+2}+(4NL-2N-2L)K^{2H-N+1}\\
&+(-2NL+2L)K^{2H-N}+4K^{H-N+1}+(4L-4)K^{H+1}-4LK^{H}\\
&+(2N-2L)K-2N+2L\Big\}. \qquad (8)
\end{aligned}
$$

4 An Optimal Depth N^*

This section seeks an optimal depth N^* of two nodes which maximizes the total shortening distance $S_H(N)$ in Eq. (8).

Let $\Delta S_H(N) \equiv S_H(N+1) - S_H(N)$, so that we have

$$
\begin{aligned}
\Delta S_H(N) = \frac{1}{(K-1)^2}\Big[&\left\{(2NL-2N)K^2+(-4NL+2N+2)K+2NL\right\}K^{2H-N-1}\\
&-4K^{H-N}+2\Big] \qquad (9)
\end{aligned}
$$

for $N = 1, 2, \ldots, H-1$.

Lemma 2 *If $N \geq 2$, then $\Delta S_H(N) < 0$.*

Proof Let $Q_H(L) \equiv \Delta S_H(N)$. Since

$$\frac{dQ_H(L)}{dL} = 2NK^{2H-N-1} > 0 \tag{10}$$

and

$$Q_H(1) = \frac{1}{(K-1)^2}\left[\left\{2(K-1)\left(\frac{K}{K-1} - N\right)\right\}K^{2H-N-1} - 4K^{H-N} + 2\right] < 0 \tag{11}$$

for $N \geq 2$, we have $Q_H(L) < 0$ for $N \geq 2$. □

Let

$$\psi = 1 - \left(\frac{1-K^{-H+1}}{K-1}\right)^2, \tag{12}$$

so that we have Lemma 3.

Lemma 3 *If $N = 1$, then we have the following:*

(i) *If $L < \psi$, then $\Delta S_H(N) < 0$.*
(ii) *If $L = \psi$, then $\Delta S_H(N) = 0$.*
(iii) *If $L > \psi$, then $\Delta S_H(N) > 0$.*

Proof (i) If $L < \psi$, then

$$S_H(2) - S_H(1) = \frac{1}{(K-1)^2}$$
$$\left[\left\{(2L-2)K^2 + (-4L+4)K + 2L\right\}K^{2H-2} - 4K^{H-1} + 2\right] < 0. \tag{13}$$

(ii) If $L = \psi$, then $S_H(2) - S_H(1) = 0$.
(iii) If $L > \psi$, then $S_H(2) - S_H(1) > 0$. □

From Lemma 2 and Lemma 3 we have the following theorem.

Theorem 4 (i) *If $L < \psi$, then $N^* = 1$.*
(ii) *If $L = \psi$, then $N^* = 1, 2$.*
(iii) *If $L > \psi$, then $N^* = 2$.*

References

1. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to algorithms* (2nd ed.). Cambridge: MIT Press.
2. Gittell, J. H. (2000). Organizing work to support relational co-ordination. *International Journal of Human Resource Management*, *11*, 517–539.
3. Robbins, S. P. (2003). *Essentials of organizational behavior* (7th ed.). Upper Saddle River: Prentice Hall.

4. Sawada, K. (2007). A model of placing a liaison in the same level of a pyramid organization structure. In *Proceedings of 2007 IEEE International Conference on Industrial Engineering and Engineering Management*, Singapore (pp. 804–806).
5. Sawada, K. (2008). Placing a liaison between two members of the same level in an organization structure of a complete binary tree. In *Proceedings of 9th ACIS International Conference on Software Engineering Artificial Intelligence, Networking, and Parallel/Distributed Computing*, Phuket, Thailand (pp. 69–72).
6. Sawada, K., & Wilson, R. (2006). Models of adding relations to an organization structure of a complete K-ary tree. *European Journal of Operational Research*, *174*, 1491–1500.
7. Takahara, Y., & Mesarovic, M. (2003). *Organization structure: Cybernetic systems foundation*. New York: Kluwer Academic/Plenum Publishers.

Clustering for Data Privacy and Classification Tasks

Klaus B. Schebesch and Ralf Stecking

Abstract Predictive classification is a part of data mining and of many related data-intensive research activities. In applications deriving from business intelligence, potentially valuable data from large databases often cannot be used in an unrestricted way. Privacy constraints may not allow the data modeler to use all of the existing feature variables in building the classification models. In certain situations, pre-processing the original data can lead to intermediate datasets, which hide private or commercially sensitive information but still contain information useful enough for building competitive classification models. To this end, we propose to cooperatively use both unsupervised Clustering and supervised Support Vector Machines. For an instance of real-life credit client scoring, we then evaluate our approach against the case of unrestricted use of all data features.

1 Introduction

The most convenient situation for predictive classification tasks which occur in many applications of business intelligence and of data mining in general is to use a large database in an unrestricted way, allowing to deploy classical supervised statistical

K. B. Schebesch (✉)
Department of Economics, "Vasile Goldiş" Western University of Arad,
310086 Arad, Romania
e-mail: kbschebesch@uvvg.ro

K. B. Schebesch
Department of Informatics, "Vasile Goldiş" Western University of Arad,
310086 Arad, Romania

R. Stecking
Department of Economics, Carl von Ossietzky University of Oldenburg,
26111 Oldenburg, Germany
e-mail: ralf.w.stecking@uni-oldenburg.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_54,

learning like LDA, LogReg or SVM in order to determine a separation functions between classes which are induced by feature variables of the data. However, in many such classification tasks there are privacy constraints. Such constraints may be such that the modeler is not allowed to use some of the existing feature variables, like e.g. race, religion, personal identification or commercially sensitive data. In general, a modeler may not be permitted to reveal the explicit content of the data altogether. Both, Clustering and SVM can address this problem in attractive ways. Both have the means to produce a restricted number of representatives for large data sets they are working on. While SVM achieves this by forwarding support vectors, clusterings return cluster representatives. Support vectors are selected representative data points describing the boundaries between classes, while the cluster centers are often averaged quantities, but they each come with a set of cluster members, which confer further information for using other representatives. In the paper we propose using Clustering and SVM cooperatively, in order to test a privacy scheme applied to empirical data sets where all informations are available but where their use is restricted. We compare the predictive performance of such restricted models to those trained on the full data. A relevant privacy-constrained situation is when a modeler cannot gain access to the full case wise data for different reasons including restricted information disclosure (for a critique of anonymization see [3]), but can nevertheless provide the compression procedure to be applied by the data owner before handing over any training data set. Here the aim is then to recommend or to configure the compression procedure while choosing a forecasting model, which can maximally exploit these compressed data. In our empirical application of credit scoring the data are highly imbalanced, i.e. there are much more non defaulting than defaulting credit clients (for a survey see [7]). A successful approach separates main behavioral classes into a number of more homogeneous subclasses by using k-means clustering [1], recasting the original problem into a multi-class learning task. The resulting multiple models then have to be combined using voting. More recently [2] introduce a Support Cluster Machine (SCM) as an extension of the standard Support Vector Machine (SVM) with RBF kernel where information about cluster sizes and cluster covariances is used by the kernel function. Experiments on large data sets show that SCM do reduce training time, leading only to slightly higher validation error when compared to full set SVM training. In the sequel we propose a highly scalable combination of standard models for clustering and classification in credit scoring which can be readily adopted by practitioners.

2 Training of SVM on Cluster Representatives

In credit scoring practice we are given a set of $N > 0$ training examples $\{x_i, y_i\}$, $i = 1, \ldots, N$, with $x_i \in \mathbb{R}^m$ the vector of m input features or attributes of credit client i (like income, age, profession, etc.) as well as the associated labels $y_i \in \{-1, 1\}$. Think of $y_i = -1$ as describing a "non-defaulting" and of $y_i = 1$ as a "defaulting" credit client, respectively. As this is past observed behavior, a sufficiently

large number N of such observed data pairs $\{x_i, y_i\}$ of clients as training examples from an unknown but sufficiently stationary generating distribution should allow the faithful estimation of a forecasting model $s(x)$, which predicts the behavior y of a new client described by feature vector x. Models $s(x)$ depend on parameters and their flexibility may be regularized by setting hyperparameters. For LDA the forecasting function is a separating hyperplane in $\mathbb{R}^m$, which is coded by an expression like $w_0^* + w_1^* x_1 + \cdots + w_j^* x_j + \cdots + w_m^* x_m$, where x_j refers to the feature j of credit clients (suppressing client index) and vector w^* is the result of optimally separating the classes by a hyperplane, making $s(x)$ depend on parameter w. Nonlinear SVM depend on parameters and hyperparameters. A popular variant of the SVM finally produces a forecasting rule (a class separating function) of the type

$$y^{pred} \;=\; \mathbf{sign}(s(x)) \;=\; \mathbf{sign}\Big(\sum_{i=1}^{N} y_i \alpha_i^* k(x_i, x) + b^*\Big),$$

with parameters $0 \le \alpha_i^* \le C$ and b^* the result of the dual SVM optimization [5]. Here x_i refers to the m-dimensional feature vector of client i, while x is the feature vector of a new client with as yet unknown defaulting behavior y. Support vectors are training examples located near the class boundaries of a SVM solution, which permit training of a classifier with the same expected out-of-sample performance as the same classifier trained on all the other training examples as well. The hyperparameter $C > 0$ controls the amount of misclassification (or "softness", potentially avoiding over-training) of the SVM model by means of which to activate the learning examples via $\alpha_i^* > 0$ in $s(x)$. Hence, C implicitly selects the effective functional form of $s(x)$. Note that $\alpha_i^* > 0$ actively contribute to the forecasting function by invoking the ith training example via a user defined kernel $k(.,.)$, for instance by the RBF kernel $k(u, v) = \exp\Big(-\sigma||u - v||^2\Big)$ for any two client feature vectors u and v, with $||.||^2$ being the Euclidean distance between u and v. The hyperparameter $\sigma > 0$ controls the locality of the function $s(x)$, that is the distance-dependent contribution of training examples to forecast the label of new clients x.

Our goal is to replace the full or "extensive" training data set $\{x_i, y_i\}, i = 1, \ldots, N$ by $n \ll N$ cluster representatives of clusters computed on that data. Such cluster representatives can assume a large variety of forms. They can be one or a few points from the cluster or some computed average over cluster members to name a few. We use clustering in order to find a shorter (compressed) description of the original training set, and, for simplicity, we also stick to non-overlapping clusterings. As a consequence, our cluster representatives should be usable as (surrogate) training points fully commensurable with the original training data as are for instance cluster centers. Hence, the kth cluster center may be computed by

$$\sum_{j \in J(k)} \lambda_j^k (x_j, y_j), \qquad \text{producing cluster center} \quad (\bar{x}, \bar{y}),$$

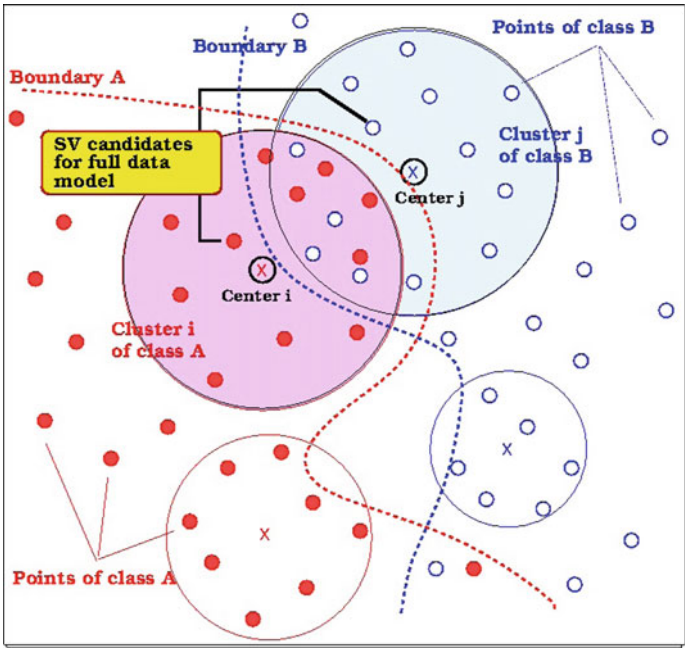


Fig. 1 Varying the number of clusters may allow even the basic clustering methods like k-means to realize via their cluster representatives an "approximation" of essential support vectors an SVM with RBF kernel would otherwise produce by training on the full data set. Clusterings of class A and class B data are produced independently, i.e. regardless of any information about the other class, respectively. Note that this is especially the case when we consider a soft margin SVM for a problem with class overlap in feature space which allows for a certain amount of misclassification (see main text)

with $\lambda_j \geq 0$ and $\sum_{j \in J(k)} \lambda_j^k = 1$. The more difficult part is of course about how to select the respective cluster member index sets $J(k)$ and the corresponding weighting schemes $\{\lambda_j^k\}$. A standard clustering procedure, which groups similar points into non-overlapping clusters and which can be used for extremely large data sets without requiring the beforehand computation of all mutual client distances is the widely accessible *k-means* algorithm. When employing clustering for data compression, using as many cluster centers as required is recommended [6]. The two (upper and lower) clustering situations from Fig. 1 may lead one to assert that one will find "intermediate" cluster numbers for which realizing the more advantageous upper situation will prevail. However, extrapolating such concepts from low dimensional intuition to higher dimensions is risky [4], hence empirical validation is in demand.

Table 1 Area under curve (AUC) statistics computed for ten randomly selected validation sets with $N = 46{,}650$ each

No. of Clusters	SVM RBF ($C = 4,\ \sigma = 2.58$) AUC (Validation Set, $N = 46{,}650$)			
	Mean	Std. Dev.	Minimum	Maximum
10	0.634	0.031	0.584	0.665
20	0.669	0.014	0.649	0.693
60	0.694	0.007	0.685	0.703
100	0.699	0.007	0.687	0.708
140	0.702	0.006	0.693	0.711
200	0.705	0.009	0.686	0.716
240	0.707	0.010	0.687	0.720
300	0.708	0.006	0.696	0.717
400	**0.712**	**0.009**	**0.699**	**0.728**
500	**0.712**	**0.008**	**0.701**	**0.722**
600	0.710	0.008	0.691	0.718
700	0.708	0.009	0.688	0.721
800	0.708	0.008	0.693	0.722
900	0.707	0.009	0.696	0.726
1,000	0.707	0.007	0.696	0.716

SVM Models with RBF kernel are trained on ten to one thousand cluster representations. The most stable cases are $n = 400$ and $n = 500$ clusters (bold-faced)

3 Empirical Results

Our data set consists of information from $N = 139{,}951$ clients of a German building and loan credit issuer. Within a time period of one year 3,692 clients refused to repay the loan. Thus, the *default rate*, based on a definition of the building and loan association, is 2.6 %. There are twelve variables per client of which eight are categorical with two up to five categories and four are quantitative. Input variables include loan related attributes like *interest rate* and *credit amount*, personal attributes like *employment status* and object related attributes like *house type*.

Our data were randomly divided into ten different training sets containing 93,301 clients and ten associated validation sets with 46,650 clients. Each training set is further subdivided into "good" and "bad" credit clients. Subsequently, standard unsupervised k–means clustering is used, partitioning the large data set into equal numbers of clusters from each class respectively, while preserving the class labels. An equal number of clusters are chosen to under-sample the much bigger class of non defaulting credit clients. However, the number of clusters necessary to best represent the training set must be determined experimentally: Too few large clusters may not include important characteristics of the data, whereas too many small clusters may inadequately highlight unpredictable regions in the data. Therefore, an ascending list of cluster numbers is tested, starting with $n = 10$ up to $n = 1{,}000$ clusters.

For each clustering, the centers (Fig. 1) are the inputs of SVM models with RBF kernel. The SVM hyperparameters for every single model from Table 1 are $\sigma = 2.58$

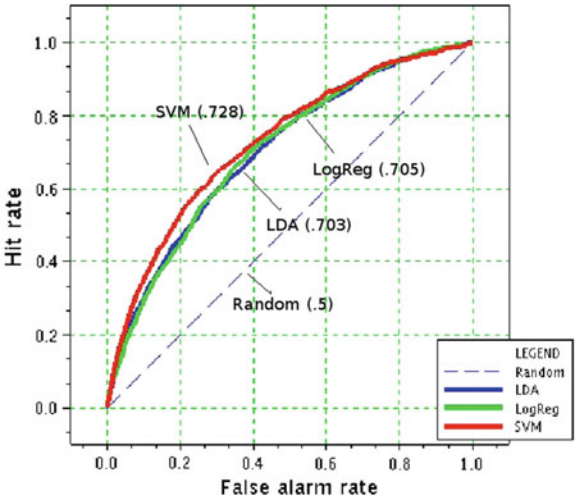


Fig. 2 ROC curves comparing out of sample prediction performance of SVM with RBF kernels trained on 400 cluster centers as shown in Table 1. Results do not differ significantly when the SVM is trained on the full data. The attained AUC is 0.728, thus RBF-SVM shows superior performance and is robust against our anonymization. When trained on the cluster centers, Linear Discriminant Analysis (LDA) and Logistic Regression (LogReg) lead to an AUC of 0.676 and 0.672 (not shown), while when trained on the full set their performance climbs significantly, to 0.703 (LDA), and 0.705 (LogReg), respectively

and $C = 4$. SVM models trained on cluster representatives can directly be used to predict individual credit client default. In order to assess the performance of the different models, *ROC curves* for all hold-out validation sets and the resulting *area under curve (AUC)* statistics are reported (for ROC and AUC see Fig. 2). Table 1 shows the mean, standard deviation, minimum and maximum of the *AUC* for ten randomly selected validation sets. The average *AUC* over the validation sets for the smallest models with just ten training examples is 0.634. The mean *AUC* is then rising continuously with growing cluster numbers, reaching a maximum of 0.712 with 400 and 500 clusters, respectively. Hereafter, the mean *AUC* decreases again, which, at least for our credit client data, confirms the assertion from the end of Sect. 2. Finally Fig. 2 also reports on the relative advantage of using RBF-SVM.

4 Conclusions

We have shown that certain privacy constraints posed on the data of classification models for credit clients can be translated into a combined clustering and classification approach which yields out-of-sample classification or forecasting performance which is comparable to models trained on the full (unconstrained) data. Other ongoing work on different credit client data sets confirms that out-of-sample performance

of k-means clustering and SVM using RBF kernels can even be at least that of SVM trained on the full data, which is certainly owing to the characteristics of these high dimensional empirical data. Hence it would be interesting to construct examples for which competitive performance of soft margin classification based on cluster representative is difficult to obtain. More specifically, one would be interested in an integrated optimization procedure which simultaneously solves or narrows down the solution alternatives for the combined clustering-classification problem.

References

1. Japkowicz, N. (2002). Supervised learning with unsupervised output separation. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 321–325).
2. Li, B., Chi, M., Fan, J., & Xue, X. (2007). Support cluster machine. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 505–512).
3. Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1710–1777.
4. Radovanović, M., Nanopulos, A., & Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, *11*, 2487–2531.
5. Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge: The MIT Press.
6. Von Luxburg, U., Williamson, R. C., & Guyon, I. (2012). Clustering: Science or art? *Workshop on Unsupervised Learning and Transfer Learning, JMLR Proceeding*, *27*, 65–79.
7. Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations*, *6*(1), 7–19.

A Decision Support Concept for Advanced Treatment Planning for Breast Cancer

Alexander Scherrer, Patrick Rüdiger, Andreas Dinges, Karl-Heinz Küfer, Ilka Schwidde and Sherko Kümmel

Abstract Breast cancer is the most common and mortal carcinosis in women and thus a major topic in clinical oncology. Treatment planning features a complex decision making about the various therapy concepts and their possible combinations. The physician plans treatment of a patient based on therapy guide lines and knowledge acquired in similar former patient cases. In particular the latter aspect requires the processing of large amounts of information in order to identify the medically relevant cases. This implies the urgent need for a decision support system in clinical routine. This work introduces a model for description of patient cases in terms of their crucial attributes and a mathematical function concept for the notion of medical relevance. These concepts are then used for an automated search on the set of former patient cases resulting in a comprehensive overview of the medically relevant ones with the therapy steps carried out therein and the observed outcomes. Provided with this information, the physician can conduct time-efficient planning of high quality breast cancer therapies for each individual patient case.

1 Introduction

Empirically acquired knowledge and expertise are of high value for complex therapy planning problems arising in clinical routine. In breast cancer therapy, physicians therefore make strong use of the experiences made in similar former patient cases for their decision making about the further treatment of a current patient case. However, searching for these medically relevant former cases among the many cases treated

A. Scherrer (✉) · P. Rüdiger · A. Dinges · K.-H. Küfer · I. Schwidde
Department of Optimization, Institut e for Industrial Mathematics (ITWM), Kaiserslautern, Germany
e-mail: alexander.scherrer@itwm.fraunhofer.de

I. Schwidde · S. Kümmel
Klinik für Senologie—Brustzentrum, Kliniken Essen-Mitte, Essen, Germany

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_55,

over the years—Kliniken Essen-Mitte face several hundred breast cancer patients per year—with their individual progress over long case histories is a very time-consuming task, which is difficult to thoroughly conduct in stressful clinical routine. This work introduces a mathematical model for breast cancer cases (Sects. 2 and 3), describes methods for the automated search of relevant former cases and their beneficial use for decision making in therapy planning (Sect. 4) and discusses the clinical benefit aspired by physicians (Sect. 5).

2 Modeling of a Patient Case

The first modeling aspect is the specification of case data with relevance for the decision making in breast cancer therapy planning. This crucial data is modeled as status attributes S_i with individual domains $\mathrm{dom}(S_i)$, whose overall number I is of the order 10^2. Examples for these attributes, are:

- *patient age* given in years;
- *menopausal status* with values pre-, peri- and postmenopausal;
- *tumor type* with more than 70 values such as invasive ductal carcinoma, invasive lobular carcinoma, ductal carcinoma in situ, flat epithelial atypia, …;
- *tumor size T* with about 60 values, where the prefix value (c, p, …) indicates the type of medical finding and the suffix assigns the actual size to groups indexed with 0, is, 1, 1mi, 1a, 1b, 1c, 2, …;
- *status of local lymph nodes N* with more than 50 values, namely a prefix analogous to the one of T and a suffix indicating the number of affected nodes with the classifications 0, 0 i+, 1 mi, 1, 1a, 1b, 1c, 2, 2a, …;
- *status of distant metastasis M* with two values, which indicate the non-existence (0) and existence (1) of metastases in combination with their location (lung, bones, liver, …);
- *estrogen receptor ER* and *progesterone receptor PR* classify the cells' degree of interference to structural changes by hormonal influence with the values + and − derived from percentage values;
- *human epidermal growth factor receptor 2 Her2* indicates the level of cell growth with the values + and − and the underlying classifications 0, 1+, 2+ and 3+;
- *Karnofsky index KI* quantifies a patient's physical performance status with values ranging from 0 % (death) to 100 % (no ailment) in steps of 10 %;
- *comorbidities* list the already experienced or existing diseases like diabetes mellitus, heart insufficiency, lung embolism, eye disease, …. The overall number of values has the order 10^2.

For more information on these and other attributes, see for example [6, 7]. These attributes are organized in a hierarchy of semantic groups. For example, T, N, M and some other attributes form the *tumor classification*, ER, PR and Her2 belong to the *immunohistochemistry*, and these and other groups and also single attributes like the

tumor type together form the semantic group *main diagnosis*. Karnofsky index and comorbidities sort into the group *secondary diagnoses*.

The status of a patient case at some point of time n is then described by a vector

$$\mathbf{s}_n = (s_{i,n})_i \in \mathbb{S} = \prod_i \mathrm{dom}(S_i)$$

containing the values of the different attributes.

The second modeling aspect is the specification of therapy components T_j with mostly similar domains $\mathrm{dom}(T_j)$, whose total number J is also of the order 10^2. The components form semantic steps in clinical routine of breast cancer and their values represent decisions about them such as rejection, approval or completion. Examples specified e.g. in [2, 3] are:

- diagnostic examinations such as *mammography, sonography of the breast* or *staging*, a combination of bone scintigraphy, chest X-ray and liver ultrasound;
- surgical treatments like *lumpectomy*, a breast surgery, or *axillary sentinel lymphectomy*, a surgery on the axillary lymph nodes;
- systemic therapies like the adjuvant (i.e. succeeding surgery) endocrine (anti-hormonal) therapy TAM $\rightarrow$ AI or the chemotherapy 4 $\times$ TC.

The history of a patient case at some point of time n is then described by a vector

$$\mathbf{t}_n = (t_{j,n})_j \in \mathbb{T} = \prod_j \mathrm{dom}(T_j)$$

Altogether, this allows for a description of a patient case

$$(\mathbf{s}_n, \mathbf{t}_n)_{n \leq N} \in (\mathbb{S} \times \mathbb{T})^{\mathbb{N}} \tag{1}$$

by means of the case status $\mathbf{s}_n$ at the various points of time $n \leq N$ and the steps $\mathbf{t}_n$ in case history connecting between them.

3 The Notion of Medical Relevance

When deciding about next therapy steps for the current patient case, a physician also relies on medically similar and thus relevant former patient cases. A former case is considered relevant, if its case status attained at some point of time deviates only slightly from the status of the current case in all attributes and the case histories leading to these two statuses are also comparable. Then the physician faces the same pre-conditions for his decision making about the next therapy steps in the current case as he was confronted with in that particular former case, which can thus provide an orientation for how to proceed for the current case. This notion of medical relevance is modeled starting from the level of attributes with functions

$$d_i : \mathrm{dom}(S_i) \times \mathrm{dom}(S_i) \longrightarrow [0, 1] \tag{2}$$

which fulfill the property of half-metrics and measure the deviation between attribute values. Exemplary functions are:

$$i = \text{tumor size T: } d_i(s_i, s_i') = \begin{cases} 0: & s_i, s_i' \text{ have suffixes from the same group} \\ & \{0\}, \{\text{mi}\}, \{1, 1\text{mi}, 1a, 1b, 1c\}, \{2\}, \ldots \\ 1: & s_i, s_i' \text{ have suffixes from different groups} \end{cases}$$

$$i = \text{Karnofsky index: } d_i(s_i, s_i') = \frac{|s_i - s_i'|}{100\ \%}$$

These functions are aggregated over the attributes and time steps to a single function

$$d_{\text{status}} : \mathbb{S}^{\mathbb{N}} \times \mathbb{S}^{\mathbb{N}} \longrightarrow [0, 1] \tag{3}$$

which measures the overall relevance of a former patient case status with respect to the current planning case. An exemplary function is

$$d_{\text{status}}\left((\mathbf{s}_n)_{n \le N}, (\mathbf{s}'_{n'})_{n' \le N'}\right) = \min_{n' \le N'} \max_i d_i\left(s_{i,N}, s'_{i,n'}\right)$$

which computes the maximum deviation over all attributes in order to ensure relevance with respect to all aspects of a case status and then takes the minimum of the obtained values over all time steps of the former case in order to identify the status most similar to the current status of the current case.

The comparison of case histories happens analogously. Medical relevance is modeled starting from the level of therapy components with half-metrics

$$d_j : \mathrm{dom}(T_j) \times \mathrm{dom}(T_j) \longrightarrow [0, 1] \tag{4}$$

Exemplary functions are

$$j = \text{staging: } d_j(t_j, t_j') = \begin{cases} 0: & t_i = t_j' \\ 1: & t_i \neq t_j' \end{cases}$$

$$j = \text{lumpectomy: } d_j(t_j, t_j') = \begin{cases} 0: & t_j' = \text{completion} \\ 1: & \text{else} \end{cases}$$

All these functions are aggregated over therapy components and time steps to a single function

$$d_{\text{history}} : \mathbb{T}^{\mathbb{N}} \times \mathbb{T}^{\mathbb{N}} \longrightarrow [0, 1] \tag{5}$$

Consider for example a former case whose status at time step N'' was identified similar to the current status of the current case. A function for comparing case histories leading to these two case statuses would be

$$d_{\text{history}}((\mathbf{t}_n)_{n\le N}, (\mathbf{t}'_{n'})_{n'\le N''}) = \max_j \min_{n\le N} \min_{n'\le N''} d_j(t_{i,n}, t'_{i,n'})$$

The expense of computing relevance values for a pair of patient cases grows linearly in the number of attributes, therapy steps and considered time steps, see also [5], and can thus be done quickly during interactive therapy planning.

The obtained function values quantify the medical relevance of a former patient case for the current decision making situation. In the terminology of automated classification, [1], functions (2) and (4) can be considered as dissimilarities and the set of all relevant former cases as a sufficiently homogeneous cluster centered around the current patient case. In the context of multi-objective decision making, [4], they can be considered as coordinate-specific distance measures on the decision space (1) and their aggregations (3) and (5) as scalarizations to an objective function.

4 The Search for Relevant Former Cases

Consider a patient case, whose current status and most recent history step in terms of status attributes and therapy components reads

$$\begin{aligned}\mathbf{t}_{N-1} &= (\text{lumpectomy} = \text{completion, axillary sentinel lymphectomy} = \text{completion}, \ldots)\\ \mathbf{s}_N &= (74 \text{ years, postmenopausal, invasive ductal carcinoma,}\\ &\quad \text{pT1c (i: 17 mm, is: 19 mm), pN0 sn- (0/2), M0,}\\ &\quad \text{ER} + (90\,\%), \text{PR} + (80\,\%), \text{Her2- (0), KI 90 \%, diabetes mellitus, } \ldots)\end{aligned}$$

This 74-year-old postmenopausal woman suffers from a carcinoma, whose origin is located in the ducts of the breast and has spread out into the surrounding tissue, see the third entry in $\mathbf{s}_N$. The previous step in case history $\mathbf{t}_{N-1}$ indicates a surgical treatment of this tumor with a lumpectomy and an axillary sentinel lymphectomy, see [3]. These surgical treatments have led to a pathologically determined (prefix p) tumor size T with a diameter of 17mm for the invasive part, which thereby falls into category 1c, and 19 mm for the non-invasive in situ part. The lymph node status N is also pathologically determined (prefix p) and shows 0 affected sentinel nodes of 2 examined ones (0/2), which yields a negative status (sn−) for the sentinel nodes and gives the suffix the value 0. There are no distant metastases (M0), estrogen receptor (ER) and progesterone receptor (PR) are positive, the human epidermal growth factor receptor (Her2) is negative, which indicates a slow cell growth, the Karnofsky index of 90 % assesses the patient to be in good physical shape and she suffers from diabetes mellitus.

This information forms the starting point for the next treatment step, in which the physician would proceed with a radiation therapy and essentially decide between the two major systemic options of an endocrine therapy like TAM → AI or some combination with a suitable chemotherapy. However, the patient age above 65 and presence of the comorbidity diabetes mellitus assign this case to a special patient

population, for which medical literature lacks evidence [7]. The knowledge acquired in former patient cases, which have at some time step featured a similar case status and from there undergone one of these therapy concepts, is thus of very high value for the physician. He/she then can take the further progression of these cases and their observed outcomes as helpful orientation for his/her decision making for the current patient case. First applying a suitable function (3) in a search run on a database with former patient cases to all pairs of the current case and former cases yields those former ones, which at some time step N'' feature a status similar to $\mathbf{s}_N$. Then applying a suitable function of the form (5) to the histories of the current case and the found ones prior to the specific time step N'' leaves the physician with the former cases that are medically relevant for the current case. He/she obtains these cases sorted in decreasing order of relevance, combined with a suitable statistic overview of their outcome.

5 Conclusions

This research work introduces a novel concept for supported treatment planning in clinical breast cancer therapy. The data model introduced in Sect. 2 allows for a uniform description of patient cases in terms of status and history. These descriptions facilitate a functional model for medical relevance based on half-metrics and suitable aggregations, see Sect. 3. These functions can be used in an automated database search, see Sect. 4, which yields all former cases similar to the current one. These search results provide helpful reference information for therapy planning in the current case in form of treatment decisions and observed progression. Altogether, this decision support concept enables the physician to plan high-quality breast cancer therapies for his/her patients in a more time-efficient and goal-oriented way.

Acknowledgments This ongoing research&development project is financed by Roche Pharma AG. The authors would like to thank Prof. Dr. Hans Hagen from the Faculty of Computer Sciences of the Technical University of Kaiserslautern (Germany), for the supervision of [5].

References

1. Bock, H. H. (1974). *Automatische Klassifikation*. Göttingen: Vandenhoeck & Ruprecht.
2. Du Bois, A., & Kümmel, S. (2013). *Standards der systemischen Therapie bei gynäkologischen Tumoren inklusive des Mammakarzinoms*. Klinik für Senologie/Brustzentrum—Kliniken Essen-Mitte: Klinik für Gynäkologie & Gynäkologische Onkologie.
3. Du Bois, A., & Kümmel, S. (2013). *Therapiestandards*. Klinik für Senologie/Brustzentrum—Kliniken Essen-Mitte: Klinik für Gynäkologie & Gynäkologische Onkologie.
4. Ehrgott, M. (2005). *Multicriteria optimization*. Berlin: Springer.
5. Rüdiger, P. (2013). Effiziente Therapieplanung bei Brustkrebs—Datenmodell, Algorithmik und Visualisierung für ein Entscheidungsunterstutzungswerkzeug. Bachelors thesis, Faculty of Computer Science, Technical University of Kaiserslautern (Germany).

6. Sobin, L. H., et al. (2009). *TNM: Classification of malignant tumours*. New York: Wiley.
7. St. Gallen Oncology Conferences. In *Proccedings of 13th international conference primary therapy of early breast cancer*, St. Gallen (Switzerland), 2013.

Comparison of Heuristics Towards Approaching a Scheduling and Capacity Planning MINLP for Hydrogen Storage in Chemical Substances

Simon Schulte Beerbühl, Magnus Fröhling and Frank Schultmann

Abstract The need for scheduling and capacity planning of electricity and energy storage technologies has risen in line with growing feed-in of intermittent wind and solar power. Hydrogen-based storage technologies usually feature non-linear as well as non-differentiable operating characteristics. This paper discusses different approaches towards integrating consumption figures derived from engineering simulations into scheduling and capacity planning problems. The comparison of best-fit functions to a heuristic-based approach of using comparably rapidly computable functions in addition with adjustment calculations shows, that reduction in complexity and calculation time leads to discrepancies at cost and furthermore at scheduling and capacity size level. Cost effects can be minimized by the heuristics, but different scheduling and capacity choice remains.

1 Introduction

Electricity storage technologies such as pumped hydro and compressed air storage have been integrated into optimization models of electricity systems, which range from regional to national levels, for example in [1] and [2]. Rasmussen et al. [3] analyzed the total demand of hydrogen as storage option on a European level. Such models, irrespective of technology, use mixed integer linear (MILP) approaches for optimization, as the electricity system models, into which the storage technologies

S. Schulte Beerbühl (✉) · M. Fröhling · F. Schultmann
Institute for Industrial Production, Karlsruhe Institute of Technology, Hertzstr. 16,
76187 Karlsruhe, Germany
e-mail: schultebeerbuehl@kit.edu

M. Fröhling
e-mail: magnus.froehling@kit.edu

F. Schultmann
e-mail: frank.schultmann@kit.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_56,

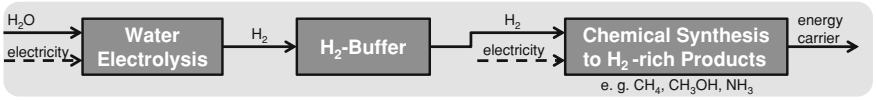


Fig. 1 Simplified flow diagram of a generalized hydrogen (H_2) energy carrier production unit

are integrated, are formulated as LP or MILP. Performance indicators such as the total efficiency have been linearized, though efficiency is usually load-dependent. This is especially true for electro- and thermo-chemical processes such as hydrogen generation by water electrolysis. Epe et al. [1], for example, introduced a piecewise linear approach to model the load-dependent efficiency and to keep the MILP environment. Since hydrogen storage, in physical or chemical way, is similar, these aspects are important for this paper. Within the scope of this paper, conversion of hydrogen to hydrogen-rich chemicals, such as methane (CH_4), methanol (CH_3OH) or ammonia (NH_3) directly after production in a water electrolysis unit (see Fig. 1) is considered.

The goal of this paper is to analyze the step of hydrogen conversion to an energy carrier (final unit in Fig. 1) and implement it into a combined scheduling and capacity planning model for determining optimal electrolysis and buffer size. We will discuss and show different approaches of integrating the synthesis unit electricity consumption into an optimization model and their influence onto the final results. Herein, these approaches shall be either formulated in a MILP or in a convex but continuous non-linear (NLP) way. Modeling hydrogen storage technologies stand-alone in a convex NLP environment, i.e. not integrated in an electricity system model, is attractive, as the hydrogen generation step contains highly non-linear but convex characteristics. Optimization runs using these approaches will be carried out for analyzing the differences in the results. A heuristic approach with a post-optimization adjustment of the consumption rate as a second step will be carried out for all simplifying approaches.

2 Methodology and Implementation

Integration of an electrolysis unit into linear or convex non-linear optimization models requires a high degree of simplification. Rasmussen et al. [3] chose an approach of constant electrolyzer efficiency in order to implement the technology into an MILP model. By focusing solely on the plant depicted in Fig. 1 within the considered planning problem and its interfaces to the electricity and energy carrier market, it is possible to refrain from the mixed integer constraints of the large electricity market models and to use a convex and differentiable NLP model instead. Consequently, the following analysis shall either result in MILP- or NLP-suitable approaches.

The objective is to maximize the annuity (EUR per year). Annualised capital costs (linearly dependent on capacity) as well as water and electricity costs will be subtracted from revenue for determining the annuity. Prices except for exogenously

Table 1 Aggregated specific electricity consumption y in kWh/1,000 Nm^3 hydrogen

Unit load (%)	100	90	80	70	60	50	40	30	25	20
Consumption y	236	235	231	228	227	228	233	245	253	263

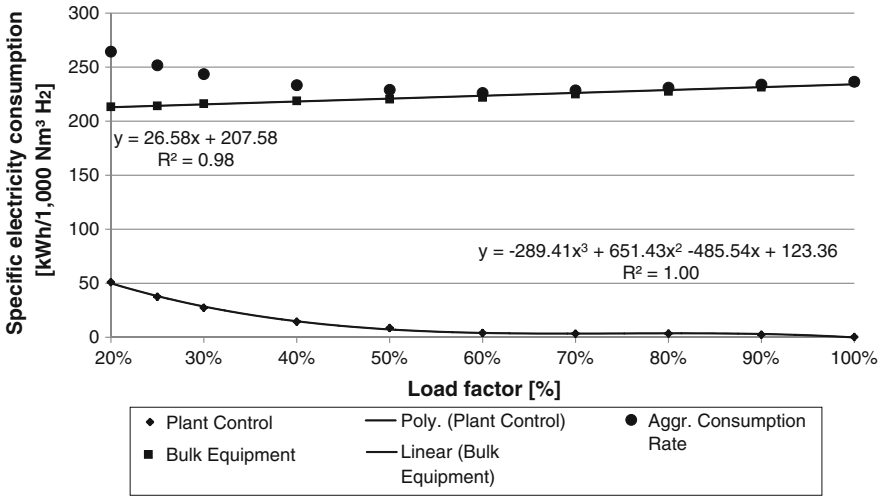


Fig. 2 Specific electricity consumption of synthesis unit

given and hourly changing electricity prices, shall be constant. The electricity consumption of the electrolysis can either be formulated in a linear (or MILP) or a non-linear way. For the synthesis unit, engineering simulations have shown, that the hydrogen-specific product yield and water consumption is constant across the load profile. However, electricity consumption of the synthesis unit changes with load (see Table 1 for details).

A detailed look into the consumption figures of individual units reveals two tendencies, which explain these values. In Fig. 2 consumption units, which are needed for thermal and pressure plant control are separated from the bulk of other consumers. By isolating these groups, we could identify a linearly decreasing specific consumption rate for the bulk consumers, which dominate total consumption. On the other side, supplementary equipment for thermal and pressure plant control is not needed at design load and only needed to a very small amount, when operating near its design load. Of course, this changes when plant load decreases to values of as low as 20–50 %. Simulation results have shown that this increase becomes apparent at approximately 50 % and increases polynomial. A cubic best fit expression of the specific consumption y per load x obtaines negligible errors. Consequently, the composite curve could be simulated ideally with a composite function, combining a linear and a cubic function. This would imply a combined expression for the objective, i.e. a MINLP.

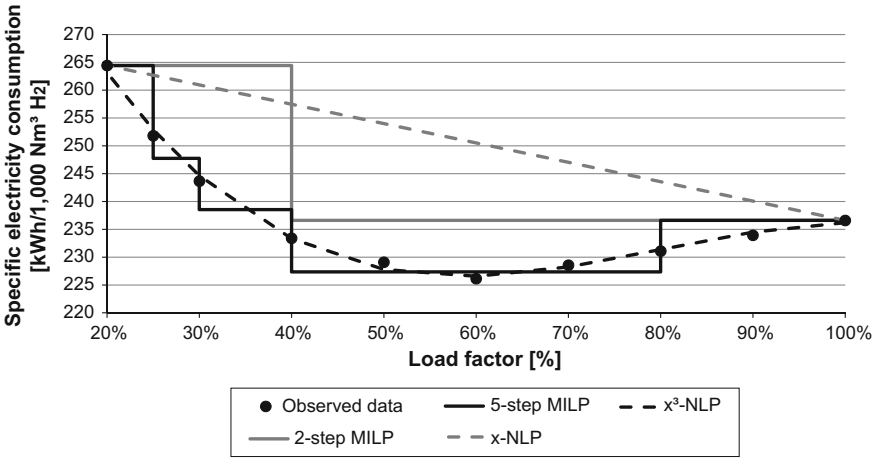


Fig. 3 Graphical illustration of approaches with regard to specific consumption

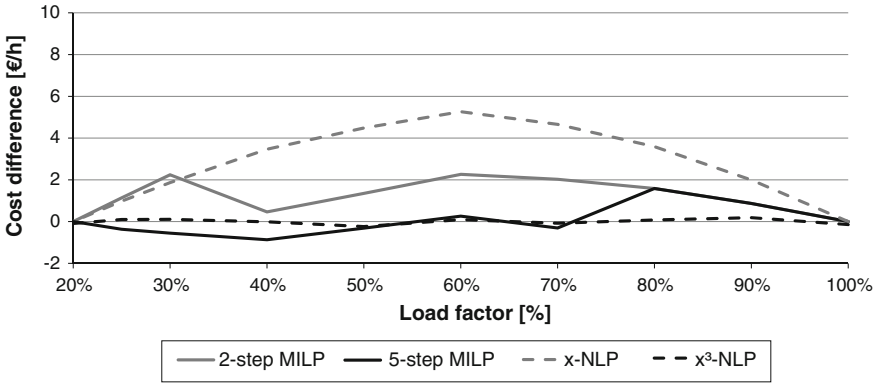


Fig. 4 Graphical illustration of approaches with regard to hourly cost differences

The tested approaches of integration reach from piecewise constant functions via a linear up to a cubic function. In Figs. 3 and 4, all solid lines represent MILP suitable approaches and all dashed lines continuous NLP suitable approaches. In all four cases, the consumption at 100 % and 20 % load shall be met, as the plant operates at either one of these points during the majority of the year and all other load stages are transit points.

Using piecewise constant specific consumption rates, as [1] did, average deviation depends upon the number of steps. Herein, we differentiate between a 5-step and a 2-step approach, the first one being more detailed and including the specific consumption minimum at 60 % load. An increase in steps leads to more binary functions and therefore more variables. In this case, the number of binary functions quadrupled and the number of variables in the reduced MILP grew by more than

Table 2 Computational characteristics of exact and simplified modeling approaches

	5-step MILP	2-step MILP	x^3-NLP	x-NLP
# Non-zeros	411,715	157,671	280,321	245,281
# Iterations	371,762	49,133	301	203
Total time elapsed	3,569 s	74 s	703 s	559 s

Intel®Core™2 Duo CPU, 2.4 GHz and 4 GB RAM, GAMS 23.9 (64 bit), using IBM ILOG CPLEX solver 12.4 resp. IPOPT 3.11

260 %, as Table 2 shows. With each step, computational time increased exponentially, which of course limits the number of steps for large problems. Figure 4 shows the hourly cost difference at the corresponding load, using an exemplary electricity price of 60 EUR per MWh.

For non-linear modeling using the IPOPT solver, the objective needs to be convex. Hourly electricity consumption, which is a component of the objective, is the product of hydrogen consumption x for each hour and specific electricity consumption y at that hour, expressed as a function of x. Therefore, $-x^2$ and $-x^4$ functions (costs have a negative sign) are convex for maximizing problems, meaning that linear or cubic functions for the specific consumption rate are acceptable. It is obvious, that the cubic function provides a better fit and incorporates the consumption minimum at 60 %. In contrast, the implemented linear function is a simple interpolation between the end points, leading to a discrepancy of approx. 10 % at medium plant loads, which translate into an overestimation of costs up to 5 EUR/hour. Comparing the problem complexity, implications are similar to the MILP case, as the problem in the linear case is reduced and therefore less time is needed.

As mentioned, the rapidly solvable approaches have been used to reduce execution time by accepting a higher degree of cost discrepancy. A post-optimization adjustment algorithm has been written for both MILP and the x-NLP approach in MATLAB. The algorithm uses the optimization results, derives actual unit load and re-calculates the electricity consumption acc. to Table 1. Hourly costs, and thus the annuity will be corrected. This procedure recovers accurate costs and reduces differences between the approaches to scheduling and capacity choice.

3 Results

The optimization problem has been solved for all four approaches, using the hourly day-ahead prices at the German electricity market EEX in 2012 as input. Deviations compared to the cubic NLP approach are presented in Table 3. It is noteworthy, that both MILP and NLP simplifications result in a closer approximation to the cubic NLP objective value than the 5-step MILP function. The reason seems to be, that—for the 5-step case—loads are only scheduled at consumption function steps, whereas in all other cases, loads are scheduled throughout the full load range from 20 % to 100 %.

Table 3 Deviations in annuity (EUR per year) of combined scheduling and capacity planning (upper part) as well as scheduling at 35 MW electrolysis (lower part) for the discussed approaches

	2-step MILP	5-step MILP	x-NLP
Combined optimization			
Electrolysis (MW)	32.369	32.302	32.369
Deviation before adjustment (EUR)	1,432	1,145	1,897
Deviation after adjustment (EUR)	487	1,347	1,028
Scheduling (35 MW)			
Deviation before adjustment (EUR)	1,457	560	1,850
Deviation after adjustment (EUR)	506	735	996

Scheduling differences in the 5-step MILP are therefore at highest and lead to the largest difference in optimum capacity as well.

The heuristic closes a major part of the objective's delta, but does not influence scheduling and capacity decisions, as it corrects costs only. Scheduling differences result from the fact that both simplified approaches do not consider the specific consumption minimum at 60 %. As the adjustment does not influence scheduling and capacity optimization, only calculation inaccuracies can be corrected. Scheduling differences are mainly resulting from the fact, that both simplified approaches do not consider the specific consumption minimum at 60 % and therefore offer a differing scheduling decision, mainly more full-load operation and consequently more hydrogen production at times of medium level electricity prices, where load reductions to 60 or 70 % are optimal. The share of scheduling- to capacity-related effect can be seen from optimization at fixed capacity (lower part of Table 3).

4 Conclusions

Consumption figures and efficiencies of chemical plants across its load profile are usually characterized by a combination of influencing factors and therefore result in complex functions, which are hardly implementable into linear or convex optimization problems. In this paper, four approaches, ranging from linear to convex parabola functions, from rather precise to more imprecise but rapid to solve functions has been compared. A heuristic post-optimization adjustment procedure can correct the cost estimates to precise results. Nevertheless, simplifying the consumption function can lead to differences in scheduling and consequently in optimal capacity, both cannot be corrected by the heuristic. The extent of difference is determined by the degree of accuracy of the simplified function and its implication to scheduling. All these aspects have to be borne in mind, when configuring faster computable optimization models for plant concepts of hydrogen storage or chemical utilization in specific and chemical plants in general.

References

1. Epe, A., Mahlke, D., Martin, A., Wagner, H.-J., Weber, C., Woll, O., et al. (2009). Betriebsoptimierung zur ökologischen Bewertung von Speichern. In R. Schultz (Ed.), *Innovative Modellierung und Optimierung von Energiesystemen* (pp. 10–13). Berlin: LIT.
2. Gollmer, R., Möller, A., Römisch, W., Schultz, R., Schwarzbach, G., & Thomas, J. (1997). Optimale blockauswahl bei der kraftwerkseinsatzplanung der VEAG. In: VDI-Berichte 1352 zur Fachtagung 'Optimierung in der Energieversorgung II', pp. 71–85. VDI, Düsseldorf.
3. Rasmussen, M., Andresen, G., & Greiner, M. (2012). Storage and balancing synergies in a fully or highly renewable pan-European power system. *Energy Policy.*, *51*, 642–651.

Influence of Fluctuating Electricity Prices due to Renewable Energies on Heat Storage Investments

Katrin Schulz, Matthias Schacht and Brigitte Werners

Abstract German electricity prices are highly influenced by the volatile and stochastic residual power load due to renewable energies. This constitutes a major challenge for energy providers, especially for municipal supply companies which provide their customers with district heat as well. Efficient and flexible combined heat and power (CHP) plants are used to fulfill the unsteady loads for heat and residual power. A heat storage offers the possibility to decouple generation from demand with respect to time. This leads to additional flexibility as it allows a power price-oriented operation of the CHP plant in order to realize profits by trading electricity on the spot market. In order to support the investment decision of a municipal energy provider, we quantify the influence of the value drivers for a heat storage which can be determined by specific demand patterns and price developments in day-to-day operations. Integrating a well-known linear model we optimize the different plant operations and power trade in a generation portfolio with and without heat storage. Results quantify the value of the storage depending on the extent and duration of fluctuations in the feed-in of renewable energies and corresponding prices.

1 Introduction

The European energy policy comprises the idea of a united European energy market as well as ambitious energy and climate policy objectives concerning the increase in energy efficiency and renewable energies and the decrease of emissions.

K. Schulz (✉) · M. Schacht · B. Werners
Faculty of Management and Economics, Chair of Operations Research and Accounting, Ruhr University Bochum, 44780 Bochum, Germany
e-mail: katrin.schulz@rub.de

M. Schacht
e-mail: matthias.schacht@rub.de

B. Werners
e-mail: or@rub.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_57,

The promotion of renewable energies is converted into German legislation by the Renewable Energies Act containing the priority feed-in of renewable energies into the grid and a specified remuneration per kilowatt hour. This highly influences electricity prices on the spot market which are market clearing prices and therefore depend on supply and demand. A large demand for power is accompanied by a high electricity price and vice versa. Due to their priority feed-in renewable energies are the first to fulfill the demand which leads to a residual load for the conventional power plants. Since the generation of renewable energies mainly depends on weather conditions, the residual load underlies high fluctuations. The resulting unsteady supply of power on the spot market induces corresponding price fluctuations.

Power supply companies are therefore faced with volatile market situations. This constitutes a challenge especially for municipal energy providers that are mainly publicly owned and have to fulfill public services, resulting in a comprehensive supply portfolio. For a detailed description of the decision situation see [6]. Within this approach we focus on municipal supply companies which provide power and district heat. Under the German law the latter is defined as heat from any source that is delivered with the help of a carrier medium [7]—water in this case.

Power and heat can be generated in a coupled process using steam to drive a turbine generating power and to feed a heat exchanger transferring heat. The so called combined heat and power (CHP) plants excel in the high utilization rate of the fuel and an increased efficiency by 10–40 % compared to separate generation which contributes to the objective to decrease emissions [3]. Thus, CHP plants are promoted by the German Co-Generation Protection Law which makes such plants even more attractive for municipal energy providers that have to deal with an increasing competitive pressure.

2 Heat and Power Generation in CHP Plants with a Heat Storage

If the supply of district heat and power belongs to the portfolio of a municipal supply company, heat and power can either be generated separately or with higher efficiency simultaneously in CHP plants. In general, a steady supply has to be guaranteed which implies that the existing facilities have to be adjusted according to the specific pattern of heat and power demand and technical capabilities. In case of power demand municipal supply companies can also buy power from the spot market or act as seller of excess power. Integrating demand and generation specifications, a linear optimization model is used to determine the optimal plant deployment. In the context of increasing competitive pressure, municipal supply companies aim at ensuring the continued supply of heat and power at minimum overall net acquisition costs (1) which are determined by the generation costs for heat and power and the difference between purchase costs and revenues from power trading [5].

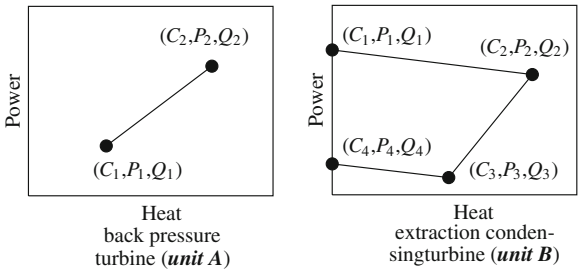


Fig. 1 In the characteristic diagram of the back pressure turbine (*unit A*) and the extraction condensing turbine (*unit B*) each extreme point E_i (with I as set of extreme points) is marked with its corresponding costs C_i, power generation P_i and heat generation Q_i

$$\min \sum_{t \in T} \left(\sum_{i \in I} C_i \cdot x_{it} + AK \cdot a_t + SP_t \cdot (h_t^+ - h_t^-) \right) \tag{1}$$

For each period t it has to be decided how the CHP plant is operated (x_{it}) and whether a start-up is necessary (a_t). The power demand can be met by own generation or purchase (h^+) and excess power can be sold (h^-) at a charge of SP_t on the spot market. The trading results are considered in the minimization objective function (1) of the optimization model as well as the hourly generation costs $\sum_{i \in I} C_i \cdot x_{it}$ and the start-up costs AK for each start-up ($a_t = 1$).

Two constraints ensure that power and heat demand are fulfilled in each period t whereby the cogeneration of heat and power can take place in two different kinds of CHP plants distinguished by the architecture of the extraction steam turbine. The basic CHP type contains a back pressure turbine which uses a constant output pressure of steam to generate heat. Therefore, power and heat are always generated at a constant ratio of power to heat and the output quantity depends on the load. The corresponding characteristic diagram of the operating points is depicted in Fig. 1 (*unit A*). A variable output ratio can be achieved with an extraction condensing turbine which possesses an extraction valve. Via this valve the extraction steam that is needed for the heat supply can be varied. The remaining steam is then conducted through a subsequent condensing steam turbine into a condenser before the two streams of steam are merged again. The resulting more flexible operating field is shown in Fig. 1 (*unit B*). According to [2] the characteristic diagrams are used to model the generation possibilities of CHP plants. The hourly power (p_t) and heat generation (q_t) is defined as a convex combination (using x_{it} for the plant operation) of the extreme points E_i [(4) and (5)]. In each period the resulting costs $\sum_{i \in I} C_i \cdot x_{it}$ (1), heat generation (q_t) and power generation (p_t) can be determined [(2) and (3)].

$$\sum_{i \in I} Q_i \cdot x_{it} = q_t \quad \forall t \in T \tag{2}$$

$$\sum_{i \in I} P_i \cdot x_{it} = p_t \quad \forall t \in T \tag{3}$$

$$\sum_{i \in I} x_{it} = 1 \quad \forall t \in T \tag{4}$$

$$x_{it} \geq 0 \quad \forall i \in I, t \in T \tag{5}$$

$$\sum_{i \in I} x_{it} = y_t \quad \forall t \in T \tag{6}$$

In order to consider the shutdown of a CHP plant a binary variable y_t is introduced with $y_t = 1$ for operation and $y_t = 0$ otherwise. Thus, (4) is replaced by (6). According to [8], y_t is used to model the start-up of a CHP plant in each period t with the binary variable a_t ($a_t = 1$ for a start-up in period t).

The presented constraints are used to optimize the plant deployment for both kinds of CHP plants which depends on the heat and power demand. The former can be predicted more or less precisely as it is mainly influenced by season and temperatures. In contrast, the power demand is volatile due to the priority feed-in of renewable energies and furthermore accompanied by fluctuating electricity prices on the spot market. Therefore, municipal energy providers face more and more frequently asynchronous demand patterns which present a major challenge concerning the plant deployment. As the generation of power and heat in CHP plants occurs in a coupled process, its flexibility is restricted depending on the kind of CHP plant (as described earlier).

In order to achieve further flexibility within the deployment planning, the investment into a heat storage with capacity $\overline{L}$ attached to the CHP plant is a strategic option. Additional income can be gained if the heat demand is covered with charged heat while the CHP plant generates excess power to sell on the spot market in times of high electricity prices. We consider a pressure accumulator that is filled with hot water at the top and cold water at the bottom of the tank which means that the discharged hot water from the top has to be replaced by cold water at the bottom immediately and vice versa [1]. Therefore, hot water can either be charged (s_t^+) or discharged (s_t^-) in each period t which is modeled in (9) with $as_t, es_t \in \{0, 1\}$ as binary variables for charging respectively discharging whereby maximum quantities ($\overline{S^+}$ and $\overline{S^-}$) have to be considered [(7) and (8)].

$$es_t \cdot \overline{S^+} \geq s_t^+ \quad \forall t \in T \tag{7}$$

$$as_t \cdot \overline{S^-} \geq s_t^- \quad \forall t \in T \tag{8}$$

$$es_t + as_t \leq 1 \quad \forall t \in T \tag{9}$$

The heat loss (V) in such a storage is assumed to be proportional with 0.05 % per hour (adapted according to [4]). The current storage level (ℓ_t) is determined by the reduced storage level of last period plus charged or minus discharged hot water.

3 Electricity Price Patterns as Value Drivers of a Heat Storage

The presented model allows to integrate a heat storage into the optimization of the plant deployment planning. Its contribution concerning flexibility depends on the existing CHP plant which is why two units namely *unit A* with a back pressure turbine and *unit B* with an extraction condensing unit are examined. In order to determine the value drivers of a heat storage with respect to fluctuating electricity prices, we analyze its use in day-to-day operations and make a comparison of the results for the same scenario with and without storage. The selected scenarios from winter months are characterized by certain demand and electricity price patterns whereby the course of the heat demand is less volatile as it is mainly temperature-dependent. Compared to that, the electricity price patterns differ clearly: While the electricity price on the spot market fluctuates slightly in scenario 1 there are volatile electricity prices with extreme upward swings to prices about €180/MWh in scenario 2. Scenario 3 shows a strong price deterioration with negative values for about 5 h. The net acquisition costs for these three scenarios are depicted in Fig. 2 whereby the costs are given in relation to the minimal costs (indicated with 100 %) for each scenario. In general it can be stated that *unit B* causes less costs due to its flexible extraction condensing turbine which allows the adaption to the electricity price. *Unit A* can only operate price-oriented if a heat storage is used to store excess heat. This is why the costs savings due to a heat storage are higher in this case compared to *unit B*. The stable development of the electricity prices in scenario 1 leads to a limited benefit of the heat storage for both units. In scenario 2 the high electricity prices make the surplus generation beneficial as additional revenues can be gained on the spot market. Although the costs saving effect is significant with a heat storage for *unit A*, the results show that *unit B* is more beneficial even without a heat storage due to its flexible power to heat ratio. In contrast, the constant heat to power ratio of *unit A* means that a high level of electricity is accompanied by a high level of heat which has to be stored. Thus, the heat storage has a high impact on the cost optimal operation for *unit A*. Scenario 3 is characterized by a strong price fall to negative prices which makes the generation of power extremely unprofitable. For the hours of negative or very low electricity prices the heat storage is used to supply the heat demand so that the CHP plant can be shut down. Therefore, the heat storage has a significant added value in this case especially for *unit A* but also for *unit B*. Figure 3 shows the high storage level shortly before negative electricity prices arise and its decrease in the corresponding hours for *unit A*. A similar effect can be observed for *unit B* in this scenario which explains the benefit of the heat storage in this case even for a flexible CHP plant as *unit B*. The added value of a heat storage depends on specific electricity price patterns and the existing CHP plant as well as the general assumption on how often the scenario will occur.

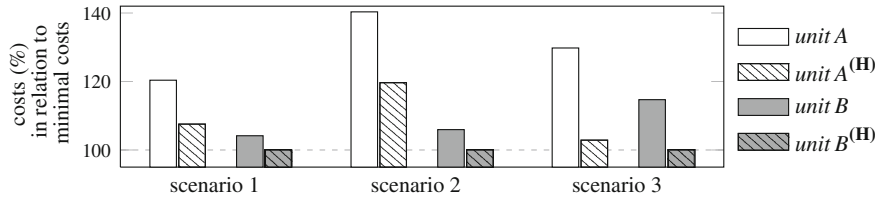


Fig. 2 Net acquisition costs in scenarios 1–3 for *unit A* and *B* with (**H**) and without heat storage

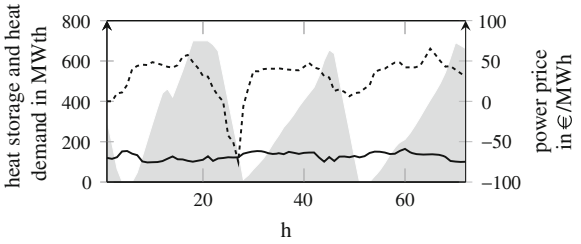


Fig. 3 Plant operation of *unit A* in scenario 2 with the given heat demand (*thick line*), power price (*dotted line*) and resulting heat storage level (*gray area*)

4 Conclusion

The priority feed-in of renewable energies leads to a volatile residual power load and corresponding prices which makes the plant deployment planning for municipal supply companies more complicated. The power and heat demand can be met with two technical divergent CHP plants which offer a different extent of flexibility. A heat storage contributes to this flexibility which is needed in order to respond well to fluctuating electricity prices in terms of an attuned trading strategy, i.e. buy (sell) power in times of low (high) electricity prices. The resulting added value of a heat storage for both kinds of CHP plants mainly depends on the particular scenario with its electricity price pattern.

We analyzed different scenarios and demonstrated exemplarily how the heat storage is used. Our results show that the benefit of the heat storage is significant with fluctuating electricity prices, especially in times of low or even negative prices as a pure generation of heat is not possible. With regard to the German objective to increase the share of renewable energies to 80 % until 2050 it can be assumed that scenarios with volatile power prices will occur more often. This will favor heat storage investments.

References

1. Fragaki, A., Andersen, A., & Toke, D. (2008). Exploration of economical sizing of gas engine and thermal store for combined heat and power plants in the UK. *Energy*, *33*(11), 1659–1670.
2. Lahdelma, R., & Hakonen, H. (2003). An efficient linear programming algorithm for combined heat and power production. *European Journal of Operational Research*, *148*(1), 141–151.
3. Mitra, S., Sun, L., & Grossman, I. E. (2013). Optimal scheduling of industrial combined heat and power plants under time-sensitive electricity prices. *Energy*, *54*, 194–211.
4. Rolfsman, B. (2004). Combined heat-and-power plants and district heating in a deregulated electricity market. *Applied Energy*, *78*(1), 37–52.
5. Rong, A., & Lahdelma, R. (2007). An efficient envelope-based Branch and Bound algorithm for non-convex combined heat and power production plants. *European Journal of Operational Research*, *183*(1), 412–431.
6. Schacht, M., & Schulz, K. (2013). Kraft-Wärme-Kopplung in kommunalen Energieversorgungsunternehmen-Volatile Einspeisung erneuerbarer Energien als Herausforderung. In: Armborst K et al. (ed) Management Science - Festschrift zum 60. Geburtstag von Brigitte Werners, Dr. Kovac, Hamburg, 337–363.
7. Topp, A. (2009). Der Begriff der Fernwärme. *Recht der Energiewirtschaft*, *4–5*, 133–138.
8. Weber, C. (2005). *Uncertainty in the electric power industry: Methods and models for decision support*. New York: Springer.

The Effects of Customer Misclassification on Cross-Training in Call Centers

Andreas Schwab and Burak Büke

Abstract The benefits of cross-training in terms of increasing responsiveness to demand fluctuations have been studied extensively in the literature. In this work, we study another important advantage of cross-training due to customer misclassification, i.e. a caller declares to face a certain problem (e.g. a hardware problem) where in fact another problem persists (e.g. a software problem). In call centers that apply no cross-training, misclassified calls need to be rerouted to agents who are able to serve the true problem, whereas cross-training enables agents to serve different problem types which reduces cycle times. We introduce two-type queueing models to study the effects of customer misclassification on cross-training in call centers. We observe that, if only a third of the agents is cross-trained, high increases in model performance can be confirmed, whereas little benefit is added by higher amounts of cross-training. We also study the effects of routing policies on cycle times.

1 Introduction

Agriculture has been forming an integral part of the economic cycle since time immemorial—today, call centers are bigger! Around 3 % of the total North American and British labor force is employed in call centers, which makes more workers than in agriculture [8]. Therefore, today's world economy is unimaginable without the call center industry. There exists a vast literature on call center operations. Akşin et al. [1]

A. Schwab (✉)
Faculty of Economics, Julius-Maximilians-Universität Würzburg, Sanderring 2,
97070 Würzburg, Germany
e-mail: schwaba83@gmail.com

B. Büke
School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building,
The King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK
e-mail: b.buke@ed.ac.uk

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_58,

and Gans et al. [4] provide an excellent introduction to this literature. Call centers are designed to handle different types of calls, and *cross-training* agents to handle multiple types of calls is a common practice in the industry [2, 3, 5]. A well-studied benefit of agent cross-training is the increase in responsiveness to demand fluctuations (see e.g. [2, 6, 7, 10]).

Another important benefit of workforce cross-training may be realized due to *customer misclassification*, where customers identify the problem they are facing wrongly, e.g. a customer declares to face a hardware problem, where in fact a software problem persists. In a call center without cross-training a customer spends a certain time at the department of her/his choice until the problem is correctly identified and s/he is rerouted to a different agent pool, where s/he goes through another problem identification phase. However, if the workers are cross-trained, problems do not need to be identified twice, which reduces the overall workload of the call center as well as the cycle times of customers. In this paper, we study cross-training policies in the light of this additional benefit. We observe that, if only a third of the agents is cross-trained, high increases in model performance can be confirmed, whereas little benefit is added by higher amounts of cross-training. This observation shows that partial cross-training outperforms full cross-training, if the cross-trained pool size is chosen with care. Moreover, the routing policies in partially cross-trained systems strongly influence customer waiting times.

2 Model Description and Assumptions

We study three basic queueing systems through simulation experiments to analyze the effects of customer misclassification on cross-training policies:

- Fully cross-trained system (`McFullCT`)
- System with no cross-training (`McNoCT`)
- Partially cross-trained system (`McM`)

The models are derivatives of the *M/M/c* queueing model with two types of calls, 0 and 1, where each type refers to a certain problem that inspires a customer to contact the call center. The customers may *misclassify* their true problems with positive probabilities, which depend on the type of problem they perceive (p_0 or p_1).

`McFullCT` indicates that the whole workforce is cross-trained, such that both types of calls can be handled by any agent in the system. Correctly classified calls experience an exponential(μ) service time, as usual in the Erlang-C model. For a misclassified call, however, we assume a *two-stage service*. Its first service time corresponds to the time an agent needs to identify the call as misclassified. We call this *problem identification* phase, what takes only a fraction (qMC_0 or qMC_1) of the regular exponential(μ) service time depending on the claimed type of the call. After this first stage, the *problem solution* phase begins with the same agent. As the employee is already familiar with the problem, the regular exponential(μ) service time decreases and only the fraction $q2nd$ is needed to serve the customer. This

fraction also depends on the type of the incoming call, i.e., $q2nd_0$ or $q2nd_1$. Since the true problem had been identified by the agent during the first service phase, the fraction $q2nd_i$ is used if the true type is i.

In McNoCT, there are equal numbers of dedicated agents for each type and no cross-trained agents. Incoming calls are therefore routed to the agent pool which serves their type. As before, correctly classified calls experience an exponential(μ) service time. A misclassified call, on the other hand, goes through the problem identification phase as in the McFullCT model, and after termination of this first service, it is rerouted to the other pool, which is able to serve its true type. If all the agents are busy—due to congestion of appropriate servers—a rule specifies the positioning of the call in its *true queue*. When an agent becomes idle, the rerouted call is handled similarly to a correctly classified call newly arriving at this pool.

McM relates to a regular *M-system*, which possesses dedicated agent pools for each type and a third cross-trained pool serving either type. This model characterizes partial cross-training; hence, it is a mixture between McFullCT and McNoCT. We need to decide whether an arriving or rerouted call shall be served by a dedicated or cross-trained agent, and this raises the *agent selection/routing problem*. By default, we assume that calls are primarily routed to dedicated agents if there exists any idle server, and if no dedicated capacity is available, the cross-trained agent pool is utilized. However, this is subject to change as part of the analysis. The *call selection/scheduling problem* is only relevant for the cross-trained agents. This pool has decide which call type will be scheduled from the queue when an agent becomes idle. By default, we assume that the call type possessing the longer queue is taken into service. We also analyze the appropriateness of this policy later. Correctly classified calls face an exponential(μ) service time as in the extreme models. However, the service rates may differ between cross-trained and type-dedicated agents. For instance, the problem identification phase may be shorter in the shared service station, as cross-training equips agents with expertise in both problem types. Misclassified calls are handled as in McFullCT or McNoCT, when they are served in the cross-trained or dedicated pool, respectively.

We assume *work-conservation* in all models, i.e. when there are simultaneously an agent idling and a call queueing, then the call is scheduled immediately. We assume a *first-come first-serve (FCFS)* scheduling policy whenever possible. Further, we fix the total number of servers in all models to 60; thus, each type is dedicated 30 in McNoCT. In McM, the same amount of dedicated servers is used for each type, and the missing number up to 60 belongs to the shared service facility. The arrival and service rates for both call types, λ_i and μ_i with $i = 0, 1$, are chosen in a way that ensures that all systems reach a steady state for all different experiment settings. These restrictions as well as the chosen rates can be seen in [9], and as base time units we use minutes.

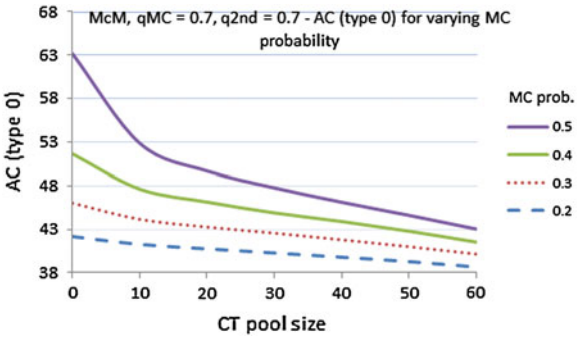


Fig. 1 Average cycle time of type 0 calls against cross-trained pool size in McM for various misclassification probabilities

3 Numerical Results

The goal of this paper is to study the benefits of cross-training for call centers particularly in the presence of customer misclassification. The main advantages of cross-training in this spirit are (1) misclassified calls do not need to be rerouted, (2) problems do not need to be identified twice, and (3) information of the problem identification phase can be further used in the problem solution phase, which reduces service time. We proceed by comparing the models in terms of the following standards[1]:

- Size of shared service station in partially cross-trained systems
- Routing policy for calls in partially cross-trained systems

3.1 Size of Shared Station in Partially Cross-Trained Systems

Based on simulation results, Fig. 1 plots the average cycle time of type 0 calls against the number of cross-trained agents between 0 and 60 in McM with $qMC = 0.7$ and $q2nd = 0.7$ for various misclassification probabilities $0.2 \leq p \leq 0.5$.

We recognize the convex shape of the relations for cross-trained pool sizes between 0 and 20, i.e. diminishing marginal utility of cross-training further agents. Thereafter, the average cycle time decreases roughly linear in the number of cross-trained agents independent of the probability of misclassification, i.e. more or less constant marginal utility. It indicates for our experiment, that cross-training up to a third of the workforce is highly efficient, as we make full use of the initial big drops in average cycle time. The other two thirds add little benefit in comparison to the first 20 cross-trained

[1] The models were compared for more standards in [9]. This paper presents an extract of the whole analysis.

agents. The linear decrease of average cycle time from cross-trained pool sizes ≥ 20 indicates that the benefits of cross-training further workers are steady independent of both the cross-trained pool size and the misclassification probability. However, the decrease of the average cycle time in the cross-trained pool size is generally stronger for higher than for lower probability of misclassification. The benchmark at one third of the workforce might be different depending on specific call center settings. Nevertheless, the value is relative to the total number of workers. Thus, we believe it to be a reasonable starting point when sizing a shared service station. We conclude that partial cross-training outperforms full cross-training, if the cross-trained pool size is chosen carefully.

3.2 Routing Policy for Calls in Partially Cross-Trained Systems

Up to this point we assumed that calls are routed to dedicated agents primarily, and only if no such capacity is available, the cross-trained agent pool is searched for idleness. If no agent is free, the call is placed in queue. However, the routing decision can also be made the other way around, i.e., that a call is preferably routed to a cross-trained agent, but if no agent is available, it goes into dedicated service. Now, we explore the impact of the latter approach on the partially cross-trained system. The rerouted calls are still primarily routed to the dedicated pool, whereas incoming calls are preferably routed to a cross-trained agent. This new approach seems particularly promising for high misclassification probabilities, since misclassified calls do not need to be rerouted, if they are served by a cross-trained agent. As of the call selection problem, we now assume that non-rerouted calls are scheduled preferably, and only if impossible, a rerouted call is taken into service. This scheduling policy matches the agent selection policy of rerouted calls. The approaches are sensible, because a rerouted call is necessarily correctly classified and is not to be rerouted again. Simulations have been performed using this setting with cross-trained pool sizes between 0 and 60.

Analog us to Fig. 1, the average cycle time for type 0 calls is plotted against cross-trained pool sizes between 0 and 60, but now with a different routing and scheduling policy. We are interested in whether the curves deviate between the two model settings. Figure 2 displays the difference in average cycle time between the two models, where a positive value indicates a smaller cycle time with the reconditioned routing policy. The new setting improves model performance in terms of average cycle time for all tested cross-trained pool sizes and misclassification probabilities; of course, with the exception of sizes of 0 (no cross-training) and 60 (full cross-training). In all cases, a higher misclassification probability is accompanied by a greater magnitude in performance difference, which approves the above-mentioned conjecture. The advantage of the new model reaches its maximum for cross-trained pool sizes of 30 or 40 depending on the misclassification probability.

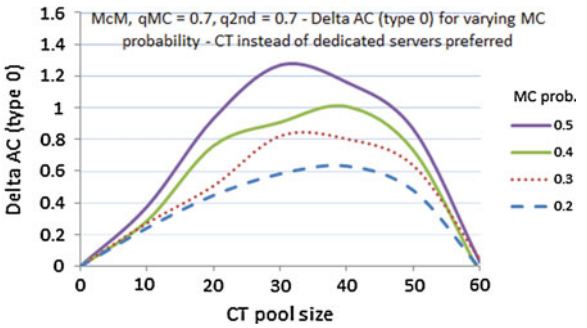


Fig. 2 Difference in average cycle time of type 0 calls against cross-trained pool size in McM between different routing policies

4 Conclusion

The essence of this paper is based on the effects of cross-training on call center performance in the face of customer misclassification. To the best of our knowledge, this field of research is largely untouched. This is dissatisfying, if we consider the benefits of workforce cross-training combined with trends towards multi-type call centers, that we believe to be a main driver of customer misclassification. We hope that this work will initiate some motivation in further research on this topic.

References

1. Akşin, O. Z., Armony, M., & Mehrotra, V. (2007). The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, *16*(6), 665–688.
2. Akşin, O. Z., Karaesmen, F., & Örmeci, E. L. (2007). A review of workforce cross-training in call centers from an operations management perspective. In D. A. Nembhard (Ed.), *Workforce Cross Training Handbook*. USA: CRC Press, Boca Raton, FL.
3. Armony M., & Maglaras C. (2001). Customer contact centers with multiple service channels. Working paper.
4. Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing and Service Operations Management*, *5*(2), 79–141.
5. Garnett, O., & Mandelbaum, A., (2000). An Introduction to Skills-Based Routing and its Operational Complexities. Teaching Note, Technion, Haifa, Israel http://ie.technion.ac.il/serveng/Lectures/SBR.pdf
6. Gurumurthi, S., & Benjaafar, S. (2004). Modeling and Analysis of Flexible Queueing Systems. *Naval Research Logistics*, *51*(5), 755–782.
7. Jordan, W. C., & Graves, S. C. (1995). Principles on the Benefits of Manufacturing Process Flexibility. *Management Science*, *41*(4), 577–594.
8. L'Ecuyer, P. (2006). Modeling and Optimization Problems in Contact Centers. *Proceedings of the Third International Conference on Quantitative Evaluation of Systems (QEST 2006)* (pp. 145–154).University of California, Riversdale: IEEE Computer Society.

9. Schwab A. (2012). Workforce cross-training in call centers. Master Thesis, University of Edinburgh, United Kingdom.
10. Tekin, E., Hopp, W. J., & Van Oyen, M. P. (2009). Pooling strategies for call center agent cross-training. *IIE Transactions*, *41*, 546–561.

Inventory Management with Transshipments Under Fill Rate Constraints

Andreas Serin and Bernd Hillebrand

Abstract Transshipments enable supply chains to reduce inventories while maintaining fill rates by sharing stored goods between different locations. In this paper, the supply chain is composed of the external manufacturer, the central warehouse and three identical retail outlets. Transshipment lead times are assumed to be negligible, while supply lead times are assumed to be deterministic as long as the sender is not out of stock. Any demand that cannot be satisfied immediately or after transshipments is lost or backlogged. A quick approximation method to estimate the expected transshipment quantities is provided. Simulation results strongly support the fit of the approximation. Numerical studies confirm the effect of lead time demand distributions on several performance measures.

1 Introduction

One approach to addressing the operating efficiency of distribution networks is to allow lateral transshipments between stocking locations at the same level (see [3]). By means of inventory pooling, stocking locations at the same echelon may reduce their safety stocks while maintaining or improving fill rates. Thus, transshipments reduce the costs of supply chain operations. The aim of this paper is to extend a single-level model according to [4] and to provide a simple method to estimate the expected transshipment quantities.

A. Serin (✉)
Production and Supply Chain Management, Mercator School of Management, Universität Duisburg-Essen, Lotharstraße 65, 47057 Duisburg, Germany
e-mail: andreas.serin@uni-due.de

B. Hillebrand
Production Management and Logistics, Supply Chain Management, Technische Universität Dortmund, Martin-Schmeißer-Weg 12, 44227 Dortmund, Germany
e-mail: bernd.hillebrand@tu-dortmund.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_59,

2 The Model

We consider a single-product two-level supply chain (One Warehouse, N Retailer) consisting of the external manufacturer, the central warehouse and three identical retail outlets under periodic review inventory management. The transshipment lead times are negligible, while the replenishment lead times are composed of deterministic shipment times and stochastic delays caused by stockouts at the central warehouse.

The incoming demand can lead to two consequences: If the pre-transshipment stock on hand exceeds the demand, the retail outlet fulfills it immediately and keeps an inventory surplus which can be offered to other retail outlets experiencing shortages. If the local demand exceeds the pre-transshipment stock on hand, the retail outlet requests an immediate lateral transshipment from the others.

Transshipments are subject to greedy policy constraints, cf. [2]. We utilize Risk Balancing Policy (RBP) equalizing the next period stockout probability for both sending or both receiving retail outlets to determine quantities to transship, cf. [4]. The remaining demand, which can't be fulfilled even by means of lateral transshipments, is backlogged or lost. At the end of each review period, every retail outlet attempts to increase its inventory position up to S_r. The central warehouse fills the orders as far as possible and raises its own inventory position up to S_c. We also utilize RBP at the central warehouse in case the central warehouse is unable to fulfill the orders completely. At the end of the period, the stock on hand is forwarded to the next period, while the backorders are backlogged or lost.

The objective function is to minimize the expected costs which are holding costs and transshipment costs.

$$\min_{S_r, S_c} \mathrm{E}C = \tau \mathrm{E}T + \eta_c \mathrm{E}I_c^+ + \sum_{i \in \mathscr{I}} \eta_r \mathrm{E}I_i^+ \tag{1}$$
$$\text{s.t. } \beta_i \geq b_r,\ i \in \mathscr{I},\ \mathscr{I} = \{1, 2, 3\}$$

We assume $\tau < \eta_c \leq \eta_r$ with respect to the unit cost parameters, and b_r denotes the desired end-customer fill rate after transshipments.

Considering the objective values from (1), we obtain the economic benefit of the transshipment policy at any particular point of the solution space:

$$\Delta \mathrm{E}C(S_r, S_c) = \tau \mathrm{E}T + \eta_c \Delta \mathrm{E}I_c^+ + 3\eta_r \Delta \mathrm{E}I_i^+ \tag{2}$$

Any transshipment flow decreases the end-of-period inventories at the retail outlets. Consequently, these outlets have to order more from the central warehouse, so the end-of-period inventories at the central warehouse are non-increasing, too. In order to minimize (1), the initial order-up-to levels S_r and S_c are pre-specified.

Let $\mathrm{E}I_i^+$ be the expected end-of-period on hand inventory, let $\mathrm{E}I_i^-$ be the expected backordered demand at the retail outlet i, and let X be the demand the retail outlet i is experiencing. Clearly, $\mathrm{E}I_i^+ - \mathrm{E}I_i^- = S_r - \mathrm{E}X$. Assuming any stationary distribution

for X, we have $\Delta \mathrm{E}I_i^+ - \Delta \mathrm{E}I_i^- = \Delta S_r$. Analogously, we conclude $\Delta \mathrm{E}I_c^+ - \Delta \mathrm{E}I_c^- = \Delta S_c - \Delta \mathrm{E}Z'$, Z' being the demand of three retail outlets addressed to the central warehouse. If we consider lost sales, we expect $|\Delta \mathrm{E}Z'| = |\Delta \mathrm{E}I_i^+|$. Otherwise, we expect $\Delta \mathrm{E}Z' = 0$.

First, let us consider $\Delta S_c = \Delta S_r = 0$. Every transshipment flow is triggered by demand which can't be fulfilled without transshipment. This demand can be satisfied only once. As every transshipment flow has exactly one source or exactly one destination, we expect $|\Delta \mathrm{E}I_i^+| = |\Delta \mathrm{E}I_i^-| = |\Delta \mathrm{E}I_c^+| = \mathrm{E}T$ to be the case, if $S_c \geq 3S_r$.

At some particular points of the solution space lying on the line $S_c = 3S_r$, we utilize analytic estimates of $\mathrm{E}I_c^+$, $\mathrm{E}I_i^+$ in case transshipments are not allowed. With an initial S_c being reasonably high and ΔS_c being sufficiently small or S_c being still increasing, we expect $\Delta \mathrm{E}I_c^- \approx 0$. Consequently, $\Delta \mathrm{E}I_c^+ \approx \Delta S_c$.

Further, we expect $|\Delta \mathrm{E}I_i^+| = |\Delta \mathrm{E}I_i^-| > \mathrm{E}T > |\Delta \mathrm{E}I_c^+|$ as a result of transshipment flows initiated to compensate the insufficient order-up-to level at the central warehouse, if $S_r \geq \mathrm{E}X$, $S_c < 3S_r$.

Unfortunately, we are not able to find out $\mathrm{E}I_i^+$ and $\mathrm{E}I_c^+$ analytically due to the limited supply from the central warehouse. Nonetheless, $\Delta \mathrm{E}C(S_r, S_c)$ is expected to be negative at any point of the solution space. As a result, the point of the solution space with the maximum transshipment quantity coincides with the minimum objective value.

The expected quantity ET to transship at time t is dependent on both S_r and S_c. For the desired end-customer fill rates $b_r = \{0.90, 0.95\}$, we expect to find minimum objective values setting $S_r \geq \mathrm{E}X$, $S_c < 3S_r$. We look at $\mathrm{E}T$ and develop an analytic approximation requiring no sophisticated computing efforts.

3 Approximation Procedure and Simulation Results

We are utilizing normal demand with parameters $\mathrm{E}X = \{200, 400, 800\}$ and $\sigma_X = 75$ as an initial point for our numerical studies. For gamma distributed demand, the corresponding parameter values resulting in the same values for $\mathrm{E}X$ and σ_X are identified. For the ease of the simulation, random demand values are rounded to the nearest integer. Negative demand values, if any, are replaced by zero.

In our approximation approach, we need to differ between the following regions of the solution space, as shown in Fig. 1. For three identical retail outlets, $S_c = 3S_r$ defines a reasonable upper bound for S_c. For a long-term view, $S_c = 3S_r$ is sufficient to establish a fill rate of 100 % at the central warehouse. Any order-up-to level $S_c > 3S_r$ would only increase the costs of the system and have no effect on $\mathrm{E}T$. The dash line represents the fill rate constraint bounding the feasible region to the bottom and to the left.

Figure 2 depicts the expected transshipment quantities per period for particular S_r and S_c values. $S_r < \mathrm{E}X$ is suppressed, as it leads to fill rates which are insufficient for any reasonable application.

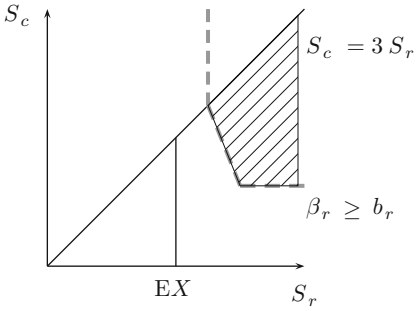


Fig. 1 Solution space

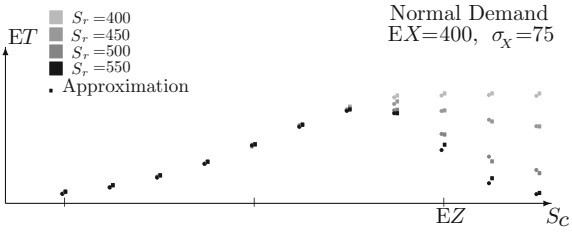


Fig. 2 Expected transshipment quantities per period

For $S_r \leq \mathrm{E}X$, $S_c < 3S_r$, the expected transshipment quantity is an increasing s-shaped curve depending on S_c independently of S_r. For $S_r > \mathrm{E}X$, $S_c < 3S_r$, it is a unimodal first increasing and then decreasing curve in S_c which is dependent on S_r, too. Above the diagonal, the expected transshipment quantity is constant in S_c depending only on S_r.

This behaviour can be explained by demand-triggered versus supply-triggered transshipment flows. Stockouts at the retail outlets can occur despite inventory positions as high as S_r. In this case, high demand triggers transshipment flows immediately. Stockouts at the central warehouse cause time-delayed transshipment flows as a consequence of the fact that retail outlets are not able to raise their inventory positions up to S_r. The interaction between the central warehouse and the retail outlets determines transshipment flows in the close neighbourhood of $S_c = \mathrm{E}Z$, if $S_c < 3S_r$ and $S_r \geq \mathrm{E}X$, Z being the threefold convolution of the demand X. For this reason, the horizontal (vertical) piece of the fill rate constraint can be approximated easily by ignoring any interdependences from the interaction between the central warehouse and the retail outlets.

For $S_r \leq \mathrm{E}X$, $S_c < 3S_r$, the expected quantities to transship are not sensitive to changes in S_r. As a result, we consider the central warehouse as the only significant factor determining $\mathrm{E}T(\cdot, S_c)$ in this part of the solution space. For the ease of computation, we assume $\frac{S_c}{3}$ to be an appropriate order-up-to level for one of three identical retail outlets.

$$\mathrm{E}T(\cdot, S_c) \approx 3 \int_{\frac{S_c}{3}}^{\infty} \left(x - \tfrac{S_c}{3}\right) dF(x) - \int_{S_c}^{\infty} (z - S_c) dF(z). \tag{3}$$

Above the diagonal, the expected quantities to transship are not sensitive to changes in S_c. These quantities are approximated in the same manner for each particular S_r value.

$$\mathrm{E}T(S_r, \cdot) \approx 3 \int_{S_r}^{\infty} (x - S_r) dF(x) - \int_{3S_r}^{\infty} (z - 3S_r) dF(z). \tag{4}$$

For $S_r > \mathrm{E}X$, $S_c < 3S_r$, we approximate $\mathrm{E}T(S_r, S_c)$ as the weighted average of (3) for the particular value of S_c and (4) for the particular value of S_r. The weights $p_n(\alpha)$ versus $1 - p_n(\alpha)$ are calculated with nth-degree polynomials of α where n is an odd number. Let $\alpha = P(Z \leq S_c)$ denote the non-stockout probability of a single stocking location serving the completely pooled demand Z, S_c being the particular order-up-to level for the periodic review policy.

Polynomials with $\lceil n/2 \rceil$ binomial coefficients perform well for S_r values up to $S_r \approx \mathrm{E}X + 2\sigma_X$. We suggest using 9th- or higher degree polynomials to improve the fit of the approximation, especially where $\mathrm{E}T(S_r, S_c)$ is still increasing in S_c for a given S_r. Though this approximation procedure doesn't need sophisticated computations, it establishes an impressive fit ($R^2 > 0.98$) for enabling reliable estimates of the expected transshipment quantities.

The solution of the entire model can be achieved by numerical methods which are beyond the scope of this paper. Herer et al. [1] describe an optimization procedure combining the advantages of simulation and stochastic optimization which can be utilized to find the minimum objective value, taking into account the relevant fill rate constraint.

4 Conclusion

Lateral transshipments lead to substantial cost benefits due to lower order-up-to levels required to establish the desired end-customer fill rate. The economic benefits depend strongly on the lead time demand distribution and unit costs under consideration. The simulation confirms cost reductions of approximately {40.55 %, 25.50 %} at the optima for $b_r = \{0.90, 0.95\}$ referring to normal demand with $\mathrm{E}X = 200$, $\sigma_X = 75$, $\eta_r = \eta_c$ and $\tau = 0.9\eta_r$. Additionally, there are some marginal improvements in terms of fill rates that the end-customers are the recipients of despite the lower order-up-to levels.

References

1. Herer, Y. T., Tzur, M., & Yücesan, E. (2006). The multilocation transshipment problem. *IIE Transactions*, *38*(3), 185–200.
2. Nonås, L., & Jörnsten, K. (2007). Optimal solutions in the multi-location inventory system with transshipments. *Journal of Mathematical Modelling and Algorithms*, *6*(1), 47–75.
3. Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research*, *210*(2), 125–136.
4. Tagaras, G. (1999). Pooling in multi-location periodic inventory distribution systems. *Omega*, *27*(1), 39–59.

Solution Method for the Inventory Distribution Problem

Takayuki Shiina

Abstract Previous research on inventory distributions between local warehouses or retailers (bases) has focused separately on either of two types of stock transshipment policies: preventive lateral transshipments or emergency lateral transshipments. Each of these has its advantages and disadvantages, and combining these policies may well enable merchandisers to achieve higher service levels. Thus, the combined use of these policies is the focus of the present study. A stochastic programming problem is formulated with demand as a stochastic variable, and the policy of using both preventive and emergency lateral transshipment is examined for its effectiveness while solution methods are examined for their efficiency.

1 Introduction

The approach to supply chain issues in recent years has been for suppliers to seek to improve service levels while satisfying a broad spectrum of consumer needs and at the same time to reduce inventory amounts and their associated expenses. However, there is a trade-off between inventory volume and service levels. To improve both at the same time, a supply chain must be carefully constructed from the planning stage, which may involve a large investment.

Lateral transshipments between retail bases are viewed as effective method for improving both inventory volume and service levels, and has come into use in some operating businesses. Two inventory transfer policies have been investigated in previous research on distribution between bases: preventive lateral transshipment [5] and emergency lateral transshipment [7]. Each has its own advantages and disadvantages, and so it is reasonable to expect that combining these will allow higher service

T. Shiina (✉)
Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, Chiba 275-0016, Japan
e-mail: shiina.takayuki@it-chiba.ac.jp

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_60,

levels to be provided. For this reason, examining the combination of these policies is the focus of the present study. Specifically, a stochastic programming problem is formulated with demand as a stochastic variable, and the policy of combined preventive and emergency lateral transshipment is examined for its effectiveness and solution methods for the formulated problem are examined for their efficiency.

2 Lateral Transshipments

In most supply chains, when a warehouse faces a stock-out situation, or when it expects a stock-out situation, it sends an order upstream. It is possible, however, that this order will have repercussions throughout the supply chain. Lateral transshipments, which regularize the risk of stock-outs by transferring inventory between bases at the retail level, are employed to reduce orders to the distribution center and improve service levels. The following two policies for lateral transshipments exist.

Preventive transshipments: Made in response to future demand expected due to inventory fluctuations prior to detecting demand increases.

Emergency transshipments: Made in response to emergencies occurring because of empty inventories, after detecting demand increases.

According to Herer, Tzur and Yucesan [3], research on problems in transferring inventory is classified into that on preventive lateral shipments, in which stock is supplied when the demand is known in advance, and that on emergency lateral shipments, in which urgent transfers are made after demand is known. Research on the former has been carried out by Karmarkar and Patel [5] and others, whereas the latter is has been studied by Tagaras [7] and others.

3 Stochastic Programming Formulation

Stochastic programming [2, 4] deals with optimization under uncertainty. A stochastic programming problem with recourse is referred to as a two-stage stochastic problem. To solve the problem, an L-shaped method [9] has been used. This approach is based on Benders [1] decomposition. The expected recourse function is piecewise linear and convex, but it is not given explicitly in advance. The L-shaped method was used to solve stochastic programs having discrete decisions in the first stage [6, 8]. The following notations are employed in the problem.

Variables

o_i	Volume of order sent to the distribution center for base i
x_{ij}	Volume of preventive lateral transshipment from base i to base j
s_i	Intended inventory volume at base i
u_i	1 if order is sent from base i to the distribution center, otherwise 0
y_{ij}^k	Volume of emergency lateral transshipment from base i to base j in scenario k

z_i^{+k} Inventory at period end at base i in scenario k
z_i^{-k} Shortage in inventory at base i in scenario k

Parameters

R_i Variable costs of orders to distribution center at base i
C_{ij} Variable costs of preventive lateral transshipment from base i to base j
S_i^0 Initial inventory at base i
E_{ij} Variable costs of emergency lateral transshipments from base i to base j
L_i Losses due to inventory outage at base i
H_i Inventory storage cost at base i
W_i Fixed order cost at base i
p^k Probability of scenario k
ξ_i^k Demand at base i in scenario k
K Total number of scenarios
I Total number of bases

The stochastic programming problem is formulated as follows.

$$\begin{aligned}
&\min \sum_{i=1}^{I} W_i u_i + \sum_{i=1}^{I} R_i o_i + \sum_{i=1}^{I}\sum_{j\neq i}^{I} C_{ij} x_{ij} + \sum_{k=1}^{K} p^k Q(s, \xi^k) \\
&\text{subject to } S_i^0 + o_i + \sum_{j=1}^{I} x_{ji} - \sum_{i=1}^{I} x_{ij} = s_i, \quad i = 1, \ldots, I \\
&o_i \leq M u_i, \quad i = 1, \ldots, I \ (M\text{: positive large number}) \\
&s_i \geq 0, o_i \geq 0, x_{ij} \geq 0, u_{ij} \in \{0, 1\}, \quad i = 1, \ldots, I, j = 1, \ldots, I, i \neq j
\end{aligned}$$

$$\begin{aligned}
Q(s, \xi^k) = \min \Bigg\{ & \sum_{i=1}^{I}\sum_{j\neq i}^{I} E_{ij} y_{ij}^k + \sum_{i=1}^{I} L_i z_i^{k-} + \sum_{i=1}^{I} H_i z_i^{k+} \Bigg| \\
& z_i^{k+} + \sum_{j=1}^{I} y_{ij}^k - (z_i^{k-} + \sum_{j=1}^{I} y_{ji}) = s_i - \xi_i^k, \quad i = 1, \ldots, I \\
& z_i^{k+}, z_i^{k-} \geq 0, y_{ij}^k \geq 0, \quad i = 1, \ldots, I, j = 1, \ldots, I, i \neq j \Bigg\}, \quad k = 1, \ldots, K
\end{aligned}$$

In the L-shaped algorithm, the following problem Master uses θ as the upper bound of the expected value for the recourse function.

$$\begin{aligned}
\text{(Master): } &\min \sum_{i=1}^{I} W_i u_i + \sum_{i=1}^{I} R_i o_i + \sum_{i=1}^{I}\sum_{j\neq i}^{I} C_{ij} x_{ij} + \theta \\
&\text{subject to } S_i^0 + o_i + \sum_{j=1}^{I} x_{ji} - \sum_{i=1}^{I} x_{ij} = s_i, \quad i = 1, \ldots, I \\
&o_i \leq M u_i, \quad i = 1, \ldots, I \\
&s_i \geq 0, o_i \geq 0, x_{ij} \geq 0, u_{ij} \in \{0, 1\}, \quad i = 1, \ldots, I, j = 1, \ldots, I, i \neq j
\end{aligned}$$

L-shaped algorithm for approximate solution

Step 1: Solve the continuous relaxation of Master, providing a solution in terms of $(\hat{u}, \hat{x}, \hat{s}, \hat{o}, \hat{\theta})$.

Step 2: Solve the second stage problem for each scenario. Because the second stage problem is feasible, the upper bound of the optimal value of the recourse function is found as $Q(\hat{s}, \xi^k), k = 1, \ldots, K$.

Step 3: If $\hat{\theta} < \sum_{k=1}^{K} p^k Q(\hat{s}, \xi^k)$, the optimality cut $\theta \geq \sum_{k=1}^{K} p^k \sum_{i=1}^{I} (s_i - \xi_i^k)\hat{\mu}_i^k$ is generated from the optimal dual solution $\hat{\mu}^k$ and added to the Master problem. Return to Step 1.

Step 4: Find the solution $(\bar{u}, \bar{x}, \bar{s}, \bar{o}, \bar{\theta})$ for the MIP problem Master. Given this solution, calculate $\sum_{i=1}^{I} W_i \bar{u}_i + \sum_{i=1}^{I} R_i \bar{o}_i + \sum_{i=1}^{I} \sum_{j \neq i}^{I} C_{ij} \bar{x_{ij}} + \sum_{k=1}^{K} p^k Q(\bar{s}, \xi^k)$ and the upper bound for the value of the optimal objective function of the original problem can be obtained.

In order to find an optimal solution with integer constraints of the original problem, the recourse function must be approximated in a feasible solution to a first stage problem satisfying the integer constraints. This must be done by solving the MIP problem Master repeatedly, and so the calculation time is potentially extremely long; however, an optimal solution is being sought for the original problem. Since the solution method shown in this paper does not necessarily approximate a recourse function completely, it provides an approximate solution for the original problem. And, it can be expected to have advantages from the viewpoint of calculation time.

4 Numerical Experiments

This experiment employed examples of lateral shipments between 20 and 25 bases. The bases were generated from a uniform distribution on a [0, 100] × [0, 100] grid. The variable cost C_{ij} of a preventive lateral transshipment from base i to base j was defined as 0.1 × (the distance between the bases), and the variable cost of an emergency lateral transshipment was defined as $E_{ij} = 1.5 \times C_{ij}$. The variable costs of orders were set at $R_i = 5$, and other parameters were set with random numbers obeying a normal distribution. Specifically, the demand at base i in scenario k, ξ_i^k, had mean 100 and variance 10; the fixed order cost at base i, W_i, had mean 200 and variance 10; the losses due to inventory outage at base i, L_i, had mean 10 and variance 1; and the inventory storage cost at base i, H_i, had mean 4 and variance 0.4.

The data sets for the different numbered scenarios (indicating problem scale) were supplied for solution by deterministic equivalent MIP conversion and by the L-shaped algorithm and the calculation times were compared. The computer used for this experiment had a 3.2 GHz Core i7-2600K (8.0 GB of memory) main processor and ran the IBM ILOG AMPL-CPLEX System 11.0 branch-and-bound solver. Both methods showed calculation times increasing with the problem scale, but the

Table 1 Results of experiment (computing time)

Base locations I	Scenarios K	L-shaped		Branch-and-bound		Relative error (%)
		Optimal objective function value	Computing time (s)	Optimal objective function value	Computing time (s)	
20	10	12,420	8	12,298	28	0.99
20	20	12,424	12	12,298	63	1.02
20	30	12,502	18	12,390	233	0.90
25	10	15,574	11	15,454	747	0.77
25	20	15,449	35	15,385	1,759	0.42
25	30	15,426	51	15,365	5,979	0.40

Table 2 Results of experiment (comparing transshipment policies)

Base locations I	Scenarios K	Variance $\mathrm{Var}[\tilde{\xi}]$	Preventive only		Emergency only		Combined policy	
			Optimal cost	Shortage ratio (%)	Optimal cost	Shortage ratio (%)	Optimal cost	Shortage ratio (%)
20	10	10	12,738	9.9	13,865	4.6	12,298	3.3
20	10	20	13,367	23.4	13,723	14.1	12,381	5.8
20	10	30	14,012	29.7	13,779	10.0	12,608	7.5
20	20	10	12,877	9.0	13,999	5.4	12,476	3.2
20	20	20	13,502	20.6	13,993	8.0	12,650	6.0
20	20	30	14,426	33.3	13,840	16.0	12,680	10.5
20	30	10	12,862	9.5	13,937	5.7	12,420	2.9
20	30	20	13,579	19.4	14,020	7.7	12,720	7.1
20	30	30	14,054	32.1	13,803	18.0	12,706	9.7

L-shaped algorithm had shorter times. As shown, solving the problem using the direct branch-and-bound algorithm for a deterministic equivalent MIP required a quite long calculation time. Thus, the L-shaped algorithm is advantageous in terms of calculation time for large-scale problems. Also, the calculation errors in this method were kept within almost 1 %, so the L-shaped method clearly provides highly accurate solutions.

Next, the difference between the costs of sending emergency and preventive lateral transshipments independently or together was compared and the effectiveness of the policy of combining emergency and preventive shipments was validated (Tables 1, 2).

For comparison with the policy of combining emergency and preventive lateral transshipments, the transfer policies restricting transshipments to either the emergency or the preventive types were reformulated, and the effectiveness of the two lateral transshipment policies was shown by comparing with the total costs of the policy of combining transshipments. The reformulation of the policy of restricting transshipments to preventive was obtained from the formulation of the policy of combining transshipments, and then eliminating the two-stage variable y_{ij}^{k}. The reformulation of the policy of restricting transshipments to emergency ship-

ments was obtained from the formulation of the policy of combining transshipments, eliminating the firts stage variable x_{ij}.

The optimal costs of the above policies and the policy of combining transshipments were compared. The numbers of demand scenarios and the standard deviations were varied in a comparison experiment. The policy of combining transshipments exhibited lower total costs than exercising policies independently, regardless of the number of scenarios or the variance. When the variance was small, the "preventive lateral transshipments only" policy had lower total costs than the "emergency lateral transshipments only" policy, and the opposite was true at high variances. This was due to the fact that the mean shortage ratio, which was defined as given below, was high when there were large fluctuations in demand. In turn, this raised shortage costs, making more emergency shipments required in order to avoid shortages.

$$\text{Mean shortage ratio}\,(\%) = \sum_{k=1}^{K} p^k \left(\sum_{i=1}^{I} z_i^{k+} / \sum_{i=1}^{I} \xi_i^k \right) \times 100. \tag{1}$$

5 Summary

In the present study, stochastic programming was employed to formulate a lateral transshipment problem, and two solution methods were examined for their efficiency in providing solutions and in combining policies enforcing preventive or emergency lateral transshipments.

The L-shaped algorithm and the direct branch-and-bound algorithm for an equivalent MIP were compared in a numerical experiment. The L-shaped algorithm was found to be advantageous in terms of calculation time for large-scale problems. It was also shown that the total costs are lowered if preventive and emergency lateral transshipment policies are combined, rather than exercising them independently.

References

1. Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, *4*, 238–252.
2. Birge, J. R., & Louveaux, F. V. (1997). *Introduction to stochastic programming*. Berlin: Springer.
3. Herer, Y. T., Tzurv, M., & Yucesan, E. (2002). Transshipments: an emerging inventory recourse to achieve supply chain leagility. *International Journal of Production Economics*, *80*, 201–212.
4. Kall, P., & Wallace, S. W. (1994). *Stochastic programming*. New York: Wiley.
5. Karmarkar, U. S., & Patel, N. (1977). The one-period N-location distribution problem. *Naval Research Logistics Quarterly*, *24*, 559–575.
6. Laporte, G., & Louveaux, F. V. (1993). The integer L-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters*, *13*, 133–142.
7. Tagaras, G. (1999). Pooling in multi-location periodic inventory distribution systems. *Omega*, *27*, 39–59.

8. Shiina, T. (2000). L-shaped decomposition method for multi-stage stochastic concentrator location problem. *Journal of the Operations Research Society of Japan*, *43*, 317–332.
9. Van Slyke, R., & Wets, R. J.-B. (1969). L-shaped linear programs with applications to optimal control and stochastic linear programs. *SIAM Journal on Applied Mathematics*, *17*, 638–663.

Application of Sampling Plan Methods: Case of Indonesian Sugar Company

Endy Suwondo, Henry Yuliando and Adi Djoko Guritno

Abstract A study of sampling plan for controling the quality of bags in a sugar company has been done. The data was taken from 25 random samples for several methods applied including *Variable Single Sampling*, *Quality Index Sampling*, and *Attribute Proportion Sampling*. It was found that the best sampling method is *Variable Single Sampling*, that gives significant result for all tested parameter and offer an efficient way for the company in doing the inspection and quality control.

1 Background

Inspection is one aspect of a quality assurance. When this inspection is carried out to accept or reject a product, then this type of inspection is called acceptance sampling. Gaspersz [5] revealed that quality assurance is an overall systematic activities implemented within the quality system and it is done regarding to the products or services to meet the requirements specified. Montgomery [12] defines quality assurance as a set of activities that ensure the level of quality of products or services are maintained correctly and resolve the quality in the view of the producers and consumers.

At present, there are a lot of reliability and accuracy test as the application of acceptance sampling in a company. Jun et al. [10] did an assesment regarding to the acceptance sampling based operating characteristics called sudden death lifetime testing. Kiermeier [11] conducted an assessment of the visualization using

E. Suwondo (✉) · H. Yuliando · A. D. Guritno
Department of Agroindustrial Technology, Gadjah Mada University, Yogyakarta, Indonesia
e-mail: endys@gadjahmada.edu

H. Yuliando
e-mail: henry@gadjahmada.edu

A. D. Guritno
e-mail: adidjoko@gadjahmada.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_61,

acceptance sampling with **R** programming language that aims to simplify user interface for acceptance sampling. Deros et al. [3] conducted a research of the acceptance sampling on three electronic products manufacturing industry in Malaysia with brackish respective case studies.

Various studies on the various types of acceptance sampling approach have been done by scientists. Chang and Hsie [2] developed a method of acceptance sampling to bridge painting quality. Graves et al. [6] evaluated the risk of producer and consumer risks in acceptance sampling with Bayesian approach. Duarte and Saraiva [4] and Jamkhaneh et al. [9] conducted an assessment of acceptance sampling method with Poisson distribution approach. Khamseh et al. [1], Hsu [7], Hsu and Hsu [8], conducted research on acceptance sampling approach that sound economically. Based on these studies, the researchers tried to apply the same acceptance sampling methods but in different areas and objects.

In this study, an acceptance sampling method applied in one of biggest Sugar Manufacturers and Refiners in Indonesia, namely PT Sweet Indolampung (as part of Sugar Group Companies), was conducted. The inspection method of white sugar packaging with a container of 50 kg sack was analyzed. Some quality attributes for each lot of sacks coming from vendors are required. In the receipt of the lot, the company conducted quality testing inspection. Here, the purpose of this study is to determine the best sampling method that can be used by the company with indicators of evaluation tools including the *probability of acceptance* (*Pa*), *the operation characteristic* (OC) curve, the *average outgoing limit* (AOL) curve, and the *average total inspection* (ATI) curve.

2 Materials and Method

The data used in this study was taken on April 12, 2012 November 30, 2012. During this period there were 14 lot orders of 50 kg sack for white sugar product. This lot further denoted as S1–S14. Quality testing conducted consist of: sack webbing test, dimension test, and inner thick test.

Data processing is done in accordance with the three different forms of sampling methods, i.e., *Variable Single Sampling* (VSS), *Quality Index Sampling* (QIS), and *Attribute-Proportion-Sampling* (APS). Phases of analysis was carried out by plotting the results of these three methods into OC curve, AOQ curve and ATI curves respectively.

VSS is a sampling method that is based on the variable product characteristics. This method considers the value of *the acceptable quality level* (*AQL*), the *rejectable quality level* (*RQL*), *producer risk* (α) and *consumer risk* (β). Here, *AQL* is the *maximum* percentage or proportion of nonconforming units in a lot that can be accepted. Whereas *RQL* is for consumer protection against bad quality lots, defined as the percentage or proportion of nonconforming units in a lot that unacceptable to the consumer. In VSS method, *W* is the criteria used to examine the sample mean ($\bar{X}$) which is used to accept or reject the lot. This is obtained by adjusting two points

between the *AQL* dan *RQL* on the OC curve. The calculation is done by using thr following equations:

$$k = \frac{Z(\alpha) \cdot \ Z(RQL) + Z(\beta) \cdot \ Z(AQL)}{Z(\alpha) + Z(\beta)} \tag{1}$$

$$n = \left(1 + \frac{k^2}{2}\right)\left(\frac{Z(\alpha) + Z(\beta)}{Z(AQL) - Z(RQL)}\right)^2 \tag{2}$$

$$W = L + k \times SD \tag{3}$$

where k is the intermediate parameter, L is specified lower limit, SD is standard deviation of sample, n is desired sample size, α and β are normal distribution value, where mostly applied at the range of 5–10 %, Z () is value in normal distribution table, and W is decision parameter.

Quality Index Sampling (QIS) is a method that estimate percentage of defect (PD) using *non central t* distribution with Q value respresents the quality index, instead of using Z value.

The *Attribute-Proportion-Sampling* Method (APS) is a method of considering the value of *AQL*, *RQL*, producer risk (α) and consumer risk (β) with the additional parameter W as the decision parameters used to examine the estimated percent defective (PD). The lot acceptance are determined by the ratio between the mean value of the samples and the W value compared to the percentage of defective (PD). If $PD \leq W$, then the lot is accepted, otherwise the lot should be rejected. Number of samples to be taken in this sampling method is obtained by using Eq. (4).

$$n = \left(\frac{Z(\alpha)\sqrt{AQL(1 - AQL)} + Z(\beta)\sqrt{RQL(1 - RQL)}}{(AQL - RQL)}\right)^2 \tag{4}$$

$$W = AQL - Z(\alpha)\sqrt{\frac{AQL(1 - AQL)}{n}} \tag{5}$$

or

$$W = RQL - Z(\beta)\sqrt{\frac{RQL(1 - RQL)}{n}}. \tag{6}$$

3 Result and Discussion

Number of samples required for the sampling plan with VSS method is 17 sheets of sacks (maximum), with a value of $k = 0.599$. By using Eq. (5), the lot acceptance probabilities are obtained at each level of quality to produce OC curve (see Fig. 1).

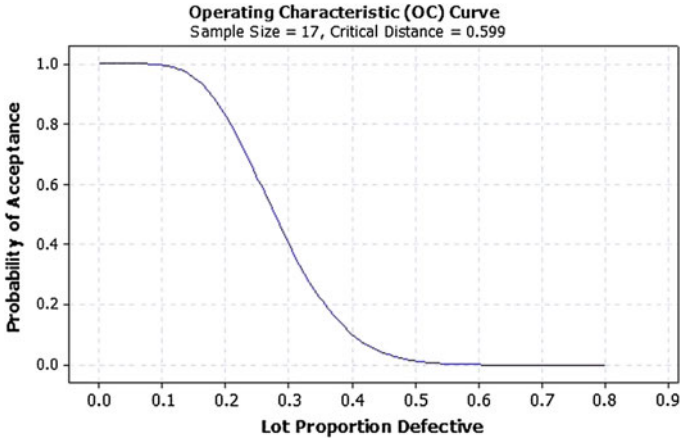


Fig. 1 OC *curve* for sampling plan with VSS

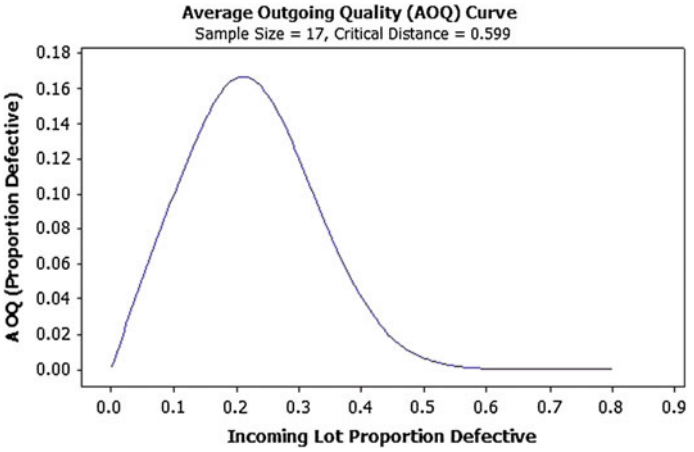


Fig. 2 AOQ *curve* for sampling plan with VSS

Figure 2 shows that the worst AOQ is reached at the level of 0.2108 or 21 % of a lot are defects. This indicates that in a lot of sugar sacks at least 80 % are in a good condition (quality). This result was based on *AQL* and *RQL* values that was employed at 0.14296 and 0.4050 or near to 15 and 40 %.

ATI curve as shown at Fig. 3 describes the average number of samples required in an inspection. It is used as evaluation tools interpret the total number of inspections in an acceptance sampling plan versus the lot fraction defective. The result for sampling plan with VSS method based on *AQL* and *RQL* values applied by the company which are 15 and 40 %, give ATI value = 703.8 for $p = 0.15$ (*AQL* = 15 %. For $p = 0.4$ (*RQL* = 40 %) the method present ATI value of 13481.4 as seen on Fig. 3.

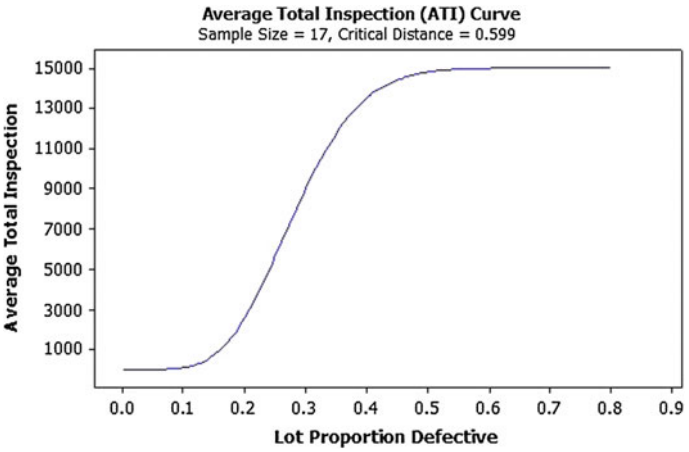


Fig. 3 ATI *curve* for sampling plan with VSS

The sampling plan method with APS revealed that to produce the OC curve takes the value of revenue and number of samples needs. Value of acceptance number c can be obtained by trial-error method that are tailored to the facing situation and the level of quality in the lot. The c values in this method is 7 and the number of samples needed as many as 24 sheets of sacks. Further, for the result of AOQ curve, APS method gave the defect fraction of $p = 0.23989$ or 24 % at AOQ $= 0.19192$. This shows that at least 81 % in a lot of good quality sacks (for AQL and RQL values are 15 and 40 % respectively). Thus with the defect fraction at 0.15 and 0.4, it is equivalent to the AOQ values as amount of 0.14687 and 0.07666.

Quality Index Sampling (QIS) method was applied by using 10 samples for each lot. The quality index is determined by the formula $Q = (U - \bar{X})/SD$ if the specified upper limit is known, or $Q = (\bar{X} - L)/SD$ when the specified lower limit known. In this study we used the specified lower limit $(L) = 0.035$. The result is presented in Table 1, showing the comparison of *PD* and *APD* against the existing of 14 lots.

The company (PT SIL) has set the amount of nonconforming level for all quality attribute of sugar sack at 20–25 % (*APD*). Therefore, when the value of $PD \geq APD$, the lot will be rejected (see Table 1). This method is quite simple as mostly implemented by the company. However, since this method ignores the value of *AQL* and *RQL*, the producer risk α and the consumer risk β as well, so that the probability of lot acceptance only determined intuitively by tester.

Finally, to select the best decision for the acceptance sampling plan method for the company it can be drawn from the tabulation of three methods employed in this study for 14 lots tested (S1–S14) as seen on Table 2.

Table 1 Decision result of 14 lots sample based on QIS method

Lot	Mean	Stdev	*Q*	*PD* (%)	*APD* (%)	Decision
S01	0.0385	0.0024	1.26	10.00	25	Accepted
S02	0.0370	0.0026	0.77	22.48	25	Accepted
S03	0.0385	0.0024	1.26	6.63	25	Accepted
S04	0.0380	0.0026	1.16	12.12	25	Accepted
S05	0.0385	0.0024	1.45	6.63	25	Accepted
S06	0.0360	0.0032	0.32	37.86	25	Rejected
S07	0.0370	0.0026	0.77	22.48	25	Accepted
S08	0.0365	0.0024	0.62	27,27	25	Rejected
S09	0.0360	0.0021	0.47	32.42	25	Rejected
S10	0.0345	0.0016	0.95	17.29	25	Accepted
S11	0.0355	0.0016	0.32	37.86	25	Rejected
S12	0.0365	0.0024	0.62	27.27	25	Rejected
S13	0.0360	0.0021	0.47	32.42	25	Rejected
S14	0.0370	0.0026	0.77	22.48	25	Accepted

Table 2 The tabulation of acceptance sampling plan methods

Methods	Current	QIS	VSS	APS
Sample amount	25	10	17	24
S01	Accepted	Accepted	Accepted	Accepted
S02	Accepted	Accepted	Accepted	Accepted
S03	Accepted	Accepted	Accepted	Accepted
S04	Accepted	Accepted	Accepted	Accepted
S05	Accepted	Accepted	Accepted	Accepted
S06	Accepted	Rejected	Rejected	Accepted
S07	Accepted	Accepted	Accepted	Accepted
S08	Accepted	Rejected	Accepted	Accepted
S09	Accepted	Rejected	Accepted	Accepted
S10	Accepted	Accepted	Rejected	Accepted
S11	Accepted	Rejected	Rejected	Accepted
S12	Accepted	Rejected	Rejected	Accepted
S13	Accepted	Rejected	Rejected	Accepted
S14	Accepted	Accepted	Accepted	Accepted

4 Conclusion

For the case of sugar company studied here, it can be concluded that based evidence showed by OC, AOQ and ATI curve for both method analyzed in this study, it was found that the best acceptance sampling plan method is Variable Single Sampling (VSS). This choice recommends the company to take 17 samples for each lot and receving 95.4 % acceptance level for each sampling

References

1. Arshadi, K. A. R., Fatemi, G. S. M. T., & Amin, N. M. (2008). Economical design of double variables acceptance sampling with inspection errors. *Journal of Faculty of Engineering*, *41*(7(109)), 959–967.
2. Chang, L., & Hsie, M. (1995). Developing acceptance-sampling methods for quality construction. *Journal of Construction Engineering and Management*, *121*(2), 246–253.
3. Deros, B. M., Peng, C. Y., Ab Rahman, M. N., Ismail, A. R., & Sulong, A. B. (2008). Assessing Acceptance Sampling Application in Manufacturing Electrical and Electronic Products. *Journal of Achievements in Materials and Manufacturing Engineering*, *31*(2), 622–628.
4. Duarte, B. P. M., Saraiva, P. M. (2008). An optimization-based approach for designing attribute acceptance sampling plans. *International Journal of Quality & Reliability Management*. doi:10.1108/02656710810898630
5. Gasperz, V. (2002). *Total Quality Management* (3rd ed.). Jakarta: PT. Gramedia Pustaka Utama.
6. Graves, S. B., Murphy, D. C., & Ringuest, J. L. (1996). Reevaluating producers and consumers risks in acceptance sampling. *Journal of Computers & Industrial Engineering*, *30*(2), 171–184.
7. Hsu, J. T. (2009). Economic design of single sample acceptance sampling plans. *Journal of Hungkuang University*, 108–122.
8. Hsu, L., Hsu, J. T. (2012). Economic design of acceptance sampling plans in a two-stage supply chain. *Advances in Decision Sciences*. doi:10.1155/2012/359082
9. Jamkhaneh, E. B., Sadeghpour-Gildeh, B., Yari, G. H. (2010). Acceptance single sampling plan by using poisson distribution. *TJMCS*, *1*(1), 6–13.
10. Jun, C. H., Balamurali, S., & Lee, S. H. (2006). Sampling plans for Weibull distributed lifetime under sudden death testing. *IEEE Transactions on Reliability*, *55*(1), 53–58.
11. Kiermeier, A. (2008). Visualising and assessing acceptance sampling plans: The R package acceptance sampling. *Journal of Statistic Software*, *26*(6).
12. Montgomery, D. C. (2005). *Introduction to Statistical Quality Control* (5th ed.). New York: Willey.

Transportation Costs and Carbon Emissions in a Vendor Managed Inventory Situation

Marcel Turkensteen and Christian Larsen

Abstract Recently, there has been much focus on carbon emissions and fuel consumption from road transport. In this paper, we consider a vendor deciding on the degree to which deliveries to geographically dispersed retailers should be consolidated. The vendor can consolidate shipments over time and deliver infrequently or deliver to many retailers simultaneously. We adapt models for determining the vendor's cost minimizing strategy and for computing emissions. We find that if the per km transportation costs increase, the vendor mainly selects smaller zones to avoid transportation, resulting in lower carbon emissions.

1 Introduction

The topic of sustainability has recently gained a great deal of attention in logistics research. Many climate scientists agree that emission of greenhouse gases such as CO_2, methane, and CFCs leads to global warming through the enhanced greenhouse effect; see [9]. Operations Research approaches can contribute to the analysis and solution of the environmental challenges; see [5]. Road transportation is an important source of emissions, responsible for about 15 % of all carbon emissions worldwide [5]. Its other negative externalities include emissions of particulate matter, noise and congestion.

Here, we consider carbon emissions resulting from joint inventory and transportation decisions. On this topic, the paper by Hoen et al. [7] discusses the relationship between inventory decisions and the choice between transport modes with different

M. Turkensteen (✉) · C. Larsen
Department of Economics and Business, Aarhus University, Fuglesangs Alle 4, 8210 Aarhus V, Denmark
e-mail: matu@asb.dk

C. Larsen
e-mail: chl@asb.dk

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_62,

carbon emission levels. The paper by Bouchery et al. [3] minimizes costs and emissions from transport and perishing of products on inventory through the selection of a delivery frequency.

In this article, we consider the relationship between inventory and routing decisions, and the resulting environmental impact in the form of carbon emissions from transportation. A vendor manages the inventory levels at retailers with stochastic demand, a set-up called *vendor managed inventory* (*VMI*), by delivering from a central warehouse. The vendor minimizes his total costs by jointly determining inventory levels at the warehouse and the retailers, and the delivery routes to be taken to replenish retailers.

We apply a model for developing the inventory and transportation policies of the vendor and combine it with a so-called engine emission model for measuring the resulting carbon emissions; see Sect. 2. In the numerical experiments in Sect. 3, we vary the per km transportation costs and analyze distribution situations with frequent and infrequent demand. Finally, the conclusions and future research directions follow in Sect. 4.

2 Computation of Optimal Policies and Carbon Emissions

In this section, we outline models for determining optimal policies for the vendor, given a certain level of per km transportation costs, and for the computation of the resulting carbon emissions.

Firstly, the vendor determines inventory levels, both at the central depot and at the retailers, and the routing decisions. Ideally, routing decisions are fully flexible. The problem is then to determine simultaneously when to visit each retailer, on which routes, and the inventory policies at each retailer. Generally, such situations are modeled as Inventory Routing Problems (IRPs), but due to the complexity of the IRP, only small instances can be solved, in particular when demand is uncertain; see e.g. [1]. Joint Replenishment Problem (JRP) approaches, on the other hand, can determine optimal inventory policies at the retailers and the depot analytically or using simulation over very long to infinite time horizons, even with uncertain demand (see e.g. [10]). However, JRP approaches tend to have a fixed cost for each delivery plus a (fixed) component for each retailer visited. In order to relate decisions to transportation costs per km and in order to compute carbon emission levels accurately, it is necessary to determine or at least estimate the route lengths through retailers on a delivery tour.

We select the JRP approach formulated in [8] that estimates the expected lengths of delivery tours using a mathematical expression from *continuous approximation*; see e.g. [4]. For this approximation to be accurate here, retailers should be more or less uniformly located across an area (the service area) and they should have identical Poisson demand distributions. For illustrative purposes, the retailers are said to be in an area with a circular shape, but this is not necessary.

In the model, the vendor can divide the service area into *zones*, where each zone contains a fixed group of retailers to which deliveries are consolidated. For each zone, the policy parameters (S, V) are determined to minimize total costs, consisting of inventory and transportation costs, but also of backorder costs. Here, S is the order-up to level of each retailer in a given zone. If more than V units demand have accumulated in the zone since the last delivery, a dispatch to the zone is made at the end of the working day. Preliminary experiments have shown that zones should be as much as possible of the same size, so to determine the optimal policy for a certain level of per km transportation costs we can simply perform exhaustive search across values of S, V, and the number of zones. When transportation costs increase, transportation can be avoided in the model by increasing V or by having small, compact zones, in both cases with lower delivery frequencies and thus higher inventory levels.

We evaluate the influence of the vendor's decisions on the expected transport-based carbon emission levels. Carbon emissions are related to distance, but also to the degree of vehicle utilization: if shipments are consolidated and vehicles are better utilized, there can both be emission savings on the distance covered by vehicles and emission increases as items are transported over longer distances. For that reason, we decompose the total average daily carbon emission into emissions depending on distance alone on the additional mass of the vehicle resulting from its load. As a consequence of this requirement off-the-shelf calculators such as the NTM calculator[1] of the Swedish Network for Transport and Environment, cannot be used. We select the engine emission model by Barth et al. [2] (pages 47–51), which is found to be the generally most accurate model in the comparison paper by Demir et al. [6]. This model relates fuel consumption and carbon emissions to several input parameters, the values of which are listed in Table 1, along with those from the chosen JRP model. The parameter values in the emission engine model are mostly as reported in [6].

We distinguish between a truck of 15 tonnes with a maximum load of 10 tonnes, and a van of 8 tonnes with a maximum load of 5 tonnes. For the van, carbon emissions per km for the empty vehicle are 0.436 kg and for transportation of an additional tonne of load 0.030 kg; if fully loaded, the load causes 32 % of the emissions. For the truck, carbon emissions per km of the empty vehicle are 0.719 kg and of each tonne of load 0.030 kg; if fully loaded, the load causes 40 % of emissions. In general, it holds that the larger the vehicle is, the smaller the share that is independent of the vehicle load. Here, the velocity is set to 50 km/h and the acceleration to 0 m/s^2.

In our numerical experiments we wish to compare carbon emissions of different solutions relative to each other, rather than absolute levels. These relative levels are determined by the division between load-based and distance-based emissions. It turns out that when different velocities and accelerations are selected, the share of load-based emissions increases with the degree of acceleration and peaks at a velocity of 50 km/h. The share lies typically between 18 and 35 % for the van and between 29 and 47 % for the truck for average accelerations smaller than 0.1 m/s^2. For larger accelerations, the share increases rapidly. Even though a constant velocity of 50 km/h may appear to be a very special case, the resulting division in emissions

[1] http://www.ntmcalc.org/Magellan/render/goodsLogistics

Table 1 Used input parameter values of the selected fuel consumption and JRP models

Symbol	Explanation	Used values
Engine emission model [2]		
m	Vehicle mass	10,000 + 1,5000 (truck)
	(load and vehicle, kg)	5,000 + 8,000 (van)
C_d	Coefficient of aerodynamic drag	0.7
A	Surface of front	3.2 (van), 5.6 (truck)
ρ	Air density (kg/m^3)	12.041
v	Velocity in m per s.	13.89 (50 km/h)
g	Gravitational constant	9.81
C_r	Rolling resistance	0.01
a	Acceleration (m/s^2)	0
η	Drive train efficiency	0.4
θ	Inclination of the road (slope)	0
JRP model [8]		
λ	Demand rate	1 per day (commodity)
W	Vehicle capacity	30 (comm.), 10 (spare p.)
T_{OP}	Open time	10 h a day
h_R	Holding cost retailer	0.5 (per hour)
h_W	Holding cost depot	0.4 (per hour)
γ_0	Fixed cost delivery tour	50
γ_1	Cost per driven km	Varies
γ_2	Cost per retailer visited	10
γ_3	Inventory cost on vehicle	1.5 (per hour)
p_0	(Stock-out cost at depot)	1,000
p_1	Stock-out cost/occurrence	5
p_2	Stock-out cost/time unit	5

is fairly typical. However, an interesting research direction is to consider emissions on trips with accelerations and decelerations, over so-called *driving cycles*.

3 Numerical Experiments

In this section, we present the most relevant of our experimental results. As the transportation costs are varied, the vendor can choose to keep more inventory to avoid transportation or to lower inventory costs by allowing for more transportation, thus changing the carbon emission levels. For a product with relatively frequent demand (a 'commodity'), one can expect that inventory levels can be changed more easily than for a product with infrequent demand (a 'spare part').

The transportation costs per km, denoted by γ_1, are varied between 10 and 300 units. We apply the JRP model with the settings of the parameter values as in Table 1 in a circular area with a radius of 7 units. In case of variations of these parameter

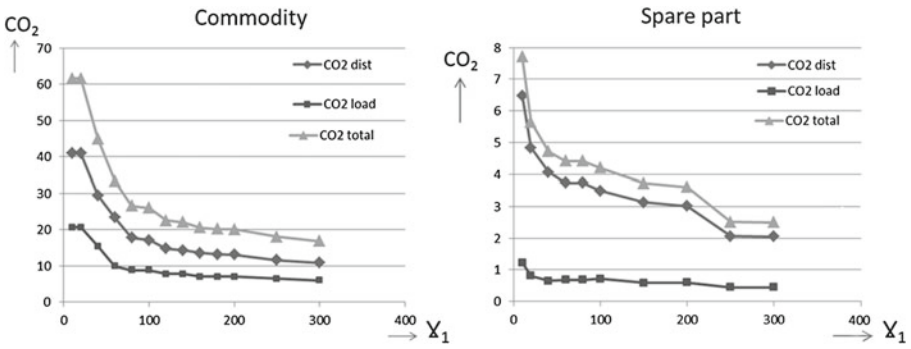


Fig. 1 CO_2 emissions (distance-based, load-based, and total) for various values of γ_1; commodity case and spare part case

values, we find that similar results are obtained as below. In the commodity case, the vehicle capacity $W = 30$ products (weighing 10 tonnes) and there are 50 retailers who each demand, on average, 1 unit per day. For the spare part case, transportation is carried out with a van with small capacity ($W = 10$, weighing 8 tonnes) and average daily demand at each of the 50 retailers is 0.1 unit. We assume that the weight of the load limits the vehicle's capacity; if space requirements limit the capacity and the load weighs only $Y\,\% < 100\,\%$ of the maximum weight, the load-based emissions are $(100 - Y)\,\%$ smaller.

Figure 1 shows that as γ_1 increases, carbon emissions decrease in the commodity case, as transportation is avoided by keeping larger inventories. Not shown is that the number of zones increases from 2 ($\gamma_1 = 20$) to 10 ($\gamma_1 = 300$) with little change in the degree of vehicle utilization. As transportation costs increase, zones become more compact and deliveries less frequent. For small intervals of γ_1, we find that the carbon emissions change shockwise: they stay at the same level for large intervals of γ_1 and change suddenly when the number of zones or, to a lesser degree, V changes.

For the spare part case, carbon emissions appear to decrease more slowly with γ_1 than for the commodity case. A possible explanation is that the vehicle utilization increases with γ_1 in the spare case part but the vehicle is almost fully utilized in all solutions to the commodity case. An increase in the number of zones reduces both load-based emissions and distance-based emissions, whereas a higher vehicle utilization reduces distance-based emissions mainly.

4 Conclusions and Future Research

In this paper, we consider a specific distribution situation from a vendor's central warehouse to dispersed retailers, where the vendor seeks minimum cost solutions. Carbon emission levels are computed for these solutions. Generally, we observe

that the amount of transportation and hence carbon emissions decrease as per km transportation costs increase and vice versa. Even though the distribution situations are different, results are similar for the spare part and commodity cases.

An interesting direction of future research is to extend our analysis to different distribution situations: with different types of vehicles, e.g. electric vehicles, retailers with different demand distributions, multiple products, and multiple warehouses. However, it may be very challenging to formulate and solve the appropriate mathematical models.

Acknowledgments This work is supported by a grant from the Nordic Council, NordForsk, project no. 25900 and entitled *Management design and evaluation of sustainable freight and logistics systems.*

References

1. Andersson, H., Hoff, A., Christiansen, M., Hasle, G., & Lkketangen, A. (2010). Industrial aspects and literature survey: Combined inventory management and routing. *Computers and Operations Research*, *37*, 1515–1536.
2. Barth M., Younglove T., Scora G. (2005). Development of a heavy-duty diesel modal emissions and fuel consumption mode. Technical report, Research report California Partners for Advanced Transit and Highways (PATH), UC Berkeley.
3. Bouchery, Y., Ghaffari, A., Jemai, Z., & Dallery, Y. (2012). Including sustainability criteria into inventory modeling. *European Journal of Operational Research*, *222*(2), 179–392.
4. Burns, L. D., Hall, R. W., Blumenfeld, D. E., & Daganzo, C. F. (1985). Distribution strategies that minimize transportation and inventory costs. *Operations Research*, *33*, 469–490.
5. Dekker, R., Bloemhof, J., & Mallidis, I. (2012). Operations research for green logistics—An overview of aspects, issues, contributions and challenges. *European Journal of Operational Research*, *219*, 671–679.
6. Demir, E., Bektas, T., & Laporte, G. (2011). A comparative analysis of several vehicle emission models for road freight transportation. *Transportation Research Part D*, *16*, 347–357.
7. Hoen, K. M. R., Tan, T., Fransoo, J. C., & Van Houtum, G. J. (2010). Effect of carbon emission regulations on transport mode selection in supply chains.Technical report, Eindhoven University of Technology.
8. Larsen C., Turkensteen M. (2013). A vendor managed inventory model using continuous approximations for route length estimates and Markov chain modeling for cost estimates. http://econ.medarbejdere.au.dk/fileadmin/Employees/Economics_Business/Diverse/Call_submissions_ActorReality.pdf.
9. Lashof, D. A., & Dilip, R. A. (1990). Relative contributions of greenhouse gas emissions to global warming. *Nature*, *344*, 529–531.
10. Viswanathan, S., & Mathur, K. (1997). ntegrated routing and inventory decisions in onewarehouse multi-retailer multi-product distribution systems. *Management Science*, *43*(3), 294–312.

Fuel Consumption Costs of Routing Uncertainty

Stephan Unger and William Cheung

Abstract We solve a car driver's routing decision problem under uncertainty in terms of fuel consumption costs. Suppose a car driver can estimate his fuel consumption for a given route between A and B. We study the optimal decision regarding which route to take, given the possibility of travelling between A and B using different routes, where each route is characterized by stochastic uncertain fuel consumption due to unknown traffic at the time of decision. We show that the cost of fuel consumption decreases significantly when taking routes with uncertain knowledge about prevailing traffic.

1 Introduction

Traffic congestion is one of the most severe cost problems faced by most businesses. A study from IBM [4] in 2010 showed that traffic congestion cost the European Union more than one percent of the gross domestic product (GDP)—or over 100 billion Euros—per year. Accordingly, U.S. drivers wasted 4.2 billion hours, 2.8 billion gallons of fuel and USD 87.2 billion due to traffic congestion in 2007. Furthermore, twenty percent of the CO_2 emissions are the byproduct of transportation. Several papers address the problem of routing and transportation optimization. Knittel [1] looks at different ways to reduce fuel consumption in transportation, and Onada [3] provides an overview of current car fuel efficiency in the largest countries.

The selected route determines how many miles per gallon a vehicle can get because this value is highly dependent on distance and traffic volume. We assume that a car

S. Unger (✉)
Department of Finance, Faculty of Busines Economics and Statistics, University of Vienna, Brünner Strasse 72, 1210 Vienna, Austria
e-mail: stephan.unger@univie.ac.at

W. Cheung
Faculty of Business Administration, Avenida Padre Tomás Pereira, Taipa, University of Macau, Macau, China
e-mail: wcheung@umac.mo

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_63,

driver is rational and tries to minimize his cost per mile. Therefore, a car driver is confronted with a decision problem as to which route to choose, knowing he has to cover a certain distance from A to B. We assume that he is in charge of the estimated fuel needed to cover this distance. What we are interested in are the costs he faces if he decides to take a different route with uncertain traffic volume and unknown distance.

In Sect. 2, we present a simple strategy that models the typical average car driver's behavior. In Sect. 3, we simulate new routings with uncertainty regarding the fuel consumption he faces to cover the new distance. The simulations are conducted with reference to lower and higher levels of traffic volume and shorter or longer distances for the new route.

The results are presented in Sect. 3.3.

2 The Model

The model for determining the sensitivities of route changes is based on the assumption that each route change is associated with a randomly generated consumption from a sample space in $\omega \in \Omega$. We start with the assumption that the typical car driver's behavior can be modeled by a type of reluctant fueling behavior. That is, we assume that each car driver only refuels his tank when it is empty, meaning that he approaches a gas station when the tank approaches empty. The random cost associated with choosing a different route is defined by the random number ε, where $\varepsilon \sim N(0, 1)$. Because we are interested in the average cost generated by choosing different routes, we need to average over all possible values of ε. Each value corresponds to a route with its own characteristics, which includes a random distance and a random traffic volume. Therefore, ε is our subject of interest. We look at different ε_j levels and their subsequent consumption costs. The various consumption costs need to be compared to various days. Each reference day starts at $t = 0$. Thus, for each month, we obtain $(t = 0, t = 1, \ldots, t = N)$ the estimation for the car driver's average monthly cost if he started with a full tank at $(t = 0, t = 1, \ldots, t = N)$. To formalize the fuel consumption behavior according to the chosen route, we give the following cost equation:

$$c_t = \mathbf{1} \sum_{n=1}^{N} \sum_{j=1}^{J} \left[\frac{\frac{X}{Y} + \varepsilon_j}{J} \right]_{t+n} = W \qquad W \cdot (F_{t+n} - F_t) \quad \text{for } t = 0, 1, 2 \ldots, n, \quad (1)$$

where W defines the tank size of the car, X defines the car driver's estimated monthly consumption, Y counts the number of days per month and F displays the fuel price. We note that the mean error of the simulated fuel consumption is zero: we do not add any drift or variation to the corresponding alternative routes.

The key in calculating the sensitivity of the fuel consumption cost to a particular routing decision lies in simulating different ε. The aggregated distances with the

given traffic volumes will lead to different refueling times for given fuel prices. For a given data set of daily observed fuel prices, we test the proposed algorithm to calculate the average cost for a car driver. Different routing decisions will lead to different refueling times, which in turn will lead to different fueling costs. We compare the savings per mile when taking a different route to the estimated cost per mile when the car driver takes the planned route.

We show that by choosing an alternative route, fueling costs per unit decrease significantly regardless of the times of refueling, knowledge about the effective distance to cover or traffic volume.

3 Routing Uncertainty

3.1 Data and Methodology

3.1.1 Data

Our data set contains daily average gasoline prices for all gas stations in Austria from 1st October 2011 to 30th June 2012. The daily gasoline data are from the official governmental e-control website for fuel price monitoring [2]. Our data set contains 271 daily summaries of gasoline prices. This sample period was chosen due to data availability. We assume that a car driver in our model has constant absolute risk aversion with an exponential utility function $U(\cdot)$.

3.1.2 Methodology

We assume that the expected total fuel used is $X = 200$ L per year. The size of the gas tank is $W = 77$ L, which is equivalent to a size of gas tank of an SUV, e.g., a Ford Explorer.

We estimate empirically the saving of consumption costs by the following steps:

Step 1. Estimate the gas used in date t, $(\frac{X}{Y} + \varepsilon)_{t+n}$, where $\varepsilon N(0, \sigma^2)$ for route a.

Step 2. Estimate the cumulative use of gas up to any given day $\sum_{n=1}^{t}(\frac{X}{Y} + \varepsilon)_{t+n}$, which is equal to the sum of gasoline used on the day in question plus the amount of gas used since the previous trip to the gas station considering different traffic conditions.

Step 3. Estimate the savings according to different traffic conditions (we assume a repeated complete fill-up of the tank).

$$c_t = \mathbf{1} \sum_{n=1}^{N} \sum_{j=1}^{J} \left[\frac{\frac{X}{Y} + \varepsilon_j}{J} \right]_{t+n} = W \qquad W \cdot (F_{t+n} - F_t) \quad \text{for } t = 0, 1, 2 \ldots, n. \quad (2)$$

3.2 Simulation of Refueling Times

Because each refueling time is associated with a different price, by simulation of different ε levels for each day of the data set, we ensure that the results are general and are not a particular random outcome due to the data structure. We can state following theorem.

Theorem 3.1 *For increasing routing uncertainty fuel costs tend to decrease.*

Proof Given W as the tank size we can define a route by it's costs c_r:

$$c_r = W(P_1 + P_2 + \cdots + P_n), \tag{3}$$

where $(P_1, P_2, \ldots, P_n)$ are the prevailing prices at gas station at the time of refueling. Pricing routing costs in advance requires estimation of

$$\mathbb{E}[c_r]. \tag{4}$$

Further we define $C(r_1, r_2, \ldots, r_n)$ as the characteristics of each route. This involves factors such as distance, traffic volume, street condition, slope, etc. which are calculated in terms of estimated fuel consumption per distance respectively it's associated fuel costs. Therefore $\frac{W}{r}$ gives us the cost for the distance a car is able to cover. When we look at a specific route we can therefore calculate its estimated cost for fueling by

$$\mathbb{E}\left[c_{r_i}\right] = \frac{W}{r_i}. \tag{5}$$

For the total estimated costs we have

$$\mathbb{E}\left[c_{avg}\right] = \frac{\mathbb{E}\left[W(P_1 + P_2 + \cdots + P_n)\right]}{\mathbb{E}\left[\frac{W}{r_1} + \frac{W}{r_2} + \cdots + \frac{W}{r_n}\right]}, \tag{6}$$

which means we are left with

$$\mathbb{E}\left[c_{avg}\right] = \frac{\mathbb{E}\left[P_1 + P_2 + \cdots + P_n\right]}{\mathbb{E}\left[\frac{1}{r_1} + \frac{1}{r_2} + \cdots + \frac{1}{r_n}\right]}. \tag{7}$$

We see that with increasing variance of the routing uncertainty, estimated costs for fueling decrease. □

One important point for achieving a valid simulation is the required condition that the estimated fuel consumption is proportional to our tank size W. Therefore, cars with bigger tanks are assumed to consume more fuel per mile than smaller cars. This assumption is necessary to ensure that our model is only dependent on one car-specific factor. The tank size is also independent of the general result (Figs. 1 and 2).

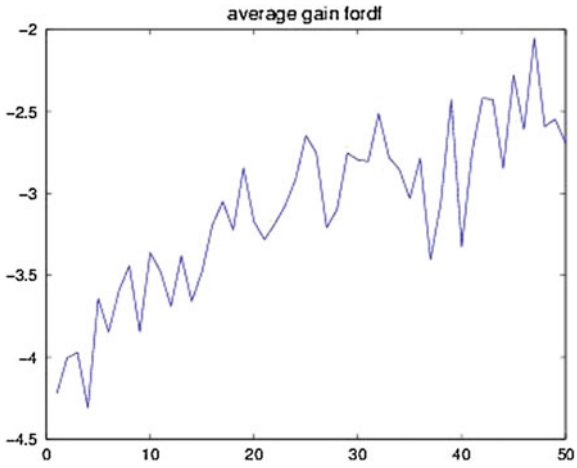


Fig. 1 Fuel costs per liter with historical prices for each σ-level for a Ford Explorer with tank size $W = 77$ L

3.3 Results

We test our proposed model on an individual car, a 2013 Ford Explorer, using historical price data as well as randomized price data for an OU process of the form

$$dS = \lambda(\mu - S)dt + \sigma dW_t, \tag{8}$$

with chosen parameters $\lambda = 0.2$, $\sigma = 0.2$, $\mu = 0.1$. We run 1,000 simulations for each day and fix the tank size for the Ford Explorer.

On average, the fueling costs tend to decrease with increasing uncertainty about the fuel consumption associated with each particular route for the Ford Explorer. The results are highly significant for the historical as well as the randomized price data. From additional testing, we can generalize our results for any random tank size. Because the simulation runs for different σs over the entire time frame, the results demonstrate independence from any possible data structure due to fuel price trends. The results imply a structural solution to car driver's routing decision problem. If a traffic jam is immanent, we know that the routing costs c_r will increase with probability one. By taking a different route the fuel costs per unit will tend to decrease. Therefore the car driver's decision problem is solved in that sense that it is never optimal to stay in the traffic jam. For the search of alternative routings the car driver is always better off.

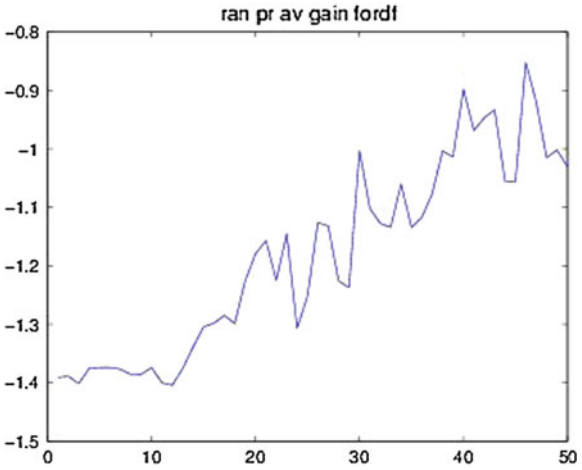


Fig. 2 Fuel costs per liter with random prizes for each σ-level for a Ford Explorer with tank size $W = 77$ L

4 Conclusion

This study demonstrates that decision-making under uncertainty can have interesting implications on optimality conditions. We examine a car driver's routing behavior under uncertain traffic conditions. A car driver who wants to drive from A to B can choose to take route a or route b. Before making his decision, he is not aware of the prevailing traffic conditions on each route. The goal of the car driver is to minimize his fuel consumption cost per unit. By assuming natural boundaries such as tank sizes and prevailing fuel prices, we determine his fuel consumption cost per unit for his chosen route. He must refuel his tank if it is empty, which serves as a counter for measuring his costs. We test our result for historical as well as randomized price data to exclude possible price dependencies. The results have the same implications for both data sets.

The key result of this study is that under certain traffic conditions, prespecified routing decisions are not optimal. Our results also indicate that changing routes increases cost savings in terms of fuel consumption per unit.

References

1. Christopher, R. (2012). Reducing petroleum consumption from transportation. *The Journal of Economic Perspectives*, *26*, 93–118.
2. E-Control (2013). $Q4$ 2011, $Q1$ 2012, $Q2$ 2012, http://www.e-control.at/de/publikationen/preistransparenz-datenbank, Fuel price monitor: Quarterly reports, Jan 2013

3. Onada, Takao. (2008). *Review of international policies for vehicle fuel efficiency*. IEA Information paper, International Energy Agency.
4. The Case for Smarter Transportation (2010). Whitepaper, September 2010.
5. U.S. Department Of Transportation. (2009). National highway traffic safety administration, average fuel economy standards passenger cars and light trucks model year 2011. *Federal Register, Rules and Regulations*, *74*(59), 255.

Optimization of Sales and Operations Planning at Shell Chemicals Europe

Thijs van Dongen and Dave van den Hurck

Abstract In the chemical industry, planning and scheduling are labor-intensive, complex, rolling processes. Interdependent decisions have to be made around different stages within the supply chain (purchases, production, distribution, exchanges, storage levels, and sales). In taking these decisions the overall enterprise margin needs to be maximized across the global supply chain. To support making these decisions the chemicals supply chain has been modeled using GMOS/NetSim, an AIMMS-based network optimization tool jointly developed by Shell Global Solutions and ORTEC. Years of extensive collaboration with various customers have made GMOS/NetSim a proven tool for strategic supply chain studies. For this project a module was developed to calculate accumulated costs/margins throughout the supply chain. The outcomes are used to do detailed margin analyses. The key challenge was to integrate the model into the monthly S&OP at SCE. Input data needs to be obtained from 15+ people around the world from various fields of expertise on a regular basis, as market conditions constantly change. Moreover, actual data is used for model validation purposes and margin analyses for past months. The key outcomes from the optimization are shared with the user community twice every month. The main benefit of this project is that we are able to establish a unified global base plan and a unified approach for fact-based decision making. The complex mathematical model behind this approach includes a great level of detail reflecting reality in everyday SCE business. This improves both the quality and speed of business decisions at 3 months (S&OP) and multi-year (business plan) horizons across the global supply chain.

T. van Dongen (✉) · D. van den Hurck
Projects and Technology, Supply Chain Optimization Software, Shell Global Solutions International BV, The Hague, The Netherlands
e-mail: Thijs.Van-Dongen@shell.com

D. van den Hurck
e-mail: Dave.Van-denHurck@shell.com

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_64,

1 Introduction

In the chemical industry, planning and scheduling are labour-intensive, complex, rolling processes. Interdependent decisions have to be made around different stages within the supply chain (purchases, production, distribution, exchanges, storage levels and sales). In taking these decisions the overall enterprise margin needs to be maximized across the global supply chain.

To support integrated decision-making the chemicals supply chain has been modelled using GMOS/NetSim, an AIMMS-based network optimization tool jointly developed by Shell Global Solutions and ORTEC. Years of extensive collaboration with various customers have made GMOS/NetSim a proven tool for strategic supply chain studies within Shell. In order to be able to do detailed margin analysis a module has been developed which provides breakdowns of the margins in the supply chain. This information helps to understand where the actual profit is made and helps in taking decisions without the necessity to re-run the model.

The key challenge is to integrate the model into the monthly S&OP processes at Shell Chemicals Europe (SCE). Input data needs to be obtained from several people around the world from various fields of expertise on a regular basis, as market conditions constantly change. Moreover, the reporting needs to be in line with expectations and the (sometimes different) business needs.

The main benefit of this project is that we are able to establish a unified global base plan and a unified approach for fact-based decision making at SCE. The complex mathematical model behind this approach includes a great level of detail reflecting reality in everyday SCE business. This improves both the quality and speed of business decisions at 3 months (S&OP) and multi-year (business plan) horizons across the global supply chain.

2 Integrated Optimization in the Chemicals Supply Chain

The goal in the chemicals business is to maximize the integrated margin while satisfying the operational constraints that are present in the supply chain. In order to achieve this goal a lot of interdependent decisions have to be made on a regular basis. These decisions have to be made on a regular basis as the market conditions constantly change introducing new opportunities and risks.

In order to be able to make decisions information is needed from various people at SCE representing different locations and different product groups. Meetings are held on a regular basis where all the stakeholders participate and the decisions are made for the coming months. These decisions are reflected in the S&OP plans for future months.

Some examples of the decisions in the chemicals supply chain are:

- The purchase of extra feedstock—It can be decided to purchase extra feedstock (if available on the market) in order to produce more product. An alternative for

this could be to use more feedstock from stock, to use more feedstock produced internally (if this is possible).
- Movement of feedstock/product—The margins constantly change and in some cases it is worthwhile to transport feedstock/product to another location to make more profit. Clearly, transportation cost (and capacity and time delay) will have to be taken into account.
- Production rates—When there is excess capacity the production of a unit with a high marginal value can sometimes be increased. In case a marginal value is negative it might be better to reduce the production rate of a production unit. If we decrease or increase the production rate operational constraints have to be taken into account (e.g. production capacity or storage). Moreover it is important to note that the production rates of productions units can be highly dependent on each other: in some cases units need to run in a fixed production ratio.
- Marginal sales—After obtaining a product we want to sell it with the highest possible margin. In order to do this we need to determine the (marginal) margin per sales channel. An alternative to selling a product can be to keep the product on storage to sell it in the future.
- Excess of feedstock—When there is an excess of feedstock available within a production facility there might be more than one unit that uses this feedstock. In this case we would like to move the feedstock to the production unit with the highest margin. An alternative might be to sell the feedstock.

In order to be able to make these decisions it is important to bring all the relevant information together. This information is needed in order to determine how SCE can make the highest margin within the operational boundaries. One of the main challenges here is to bring all the data together in such a way that the same assumptions are used everywhere. Below we provide some examples of information that is needed to be able to make optimal decisions across the supply chain:

- Sales forecasts—How much product can you sell in each location and at which price?
- Purchase forecasts—How much feedstock can you purchase in each location at which price?
- Cost forecasts—How much will your (other) costs be (e.g. transportation or production cost)?
- Capacity forecasts—What will the capacity of your production units be? Is maintenance planned next month?
- Storage forecasts—What is the storage capacity in the coming months?

Once the data have been obtained it remains challenging to create high-quality plans for the different locations and product groups as there are various dependencies and boundaries that need to be taken into account. An example of such a dependency is a production unit that produces more than one product in a fixed ratio. One of these products may have a very high margin, but one will be stuck with the low margin product. It can be very time consuming to make such plans as the market conditions

constantly change. Moreover, the plans for the different regions and product groups need to be consistent with each other and preferably created within a unified approach.

In order to support decision-making and the quality of the plans the chemicals supply chain has been modelled with the use of GMOS/NetSim. With a model of the integrated supply chain it becomes a lot easier to make integrated optimal decisions within reasonable time limits. Moreover, with the help of an integrated model it is only a small step towards scenario analysis to assess the impact of the various uncertainties in the supply chain.

3 Supply Chain Optimization with GMOS/NetSim

GMOS/NetSim is an AIMMS-based network optimization tool that has been jointly developed by Shell Global Solutions and ORTEC. The tool is used within various fields within Shell and is mostly used to do supply chain studies on a strategic or tactical level.

GMOS/NetSim has been set up in such a way that the subject matter expert (SME) can set up a network without extensive knowledge of the mathematical techniques behind it. A dedicated team provides support and training such that the potential of the tool is fully utilized by the SME. Moreover, project specific adjustments to the model or user interface can be made in case this is required. In some cases the SME is assisted by the GMOS/NetSim team to set up the model.

The underlying mathematical model is a MINLP which may become a MIP, NLP or LP when parts of the mathematical model are not used. Only those parts (constraints and/or variables) of the model are activated (or generated) for which the user specifies the input parameters. Depending on the problem type the model is solved with CPLEX (MIP/LP), CONOPT (NLP) or AOA (MINLP). In the case of SCE we solve an NLP with CONOPT consisting of around 300.000 constraints and 300.000 variables.

3.1 Building Blocks

Like in most network models the main principle behind GMOS/NetSim is that the sum of the flows that are used should be equal to the sum of the flows that become available. In simple network models this usually means that the sum of transport into a node should be equal to the sum of transport out of node. In GMOS/NetSim there are multiple ways for flows to become available or to be used (the building blocks). Due to this assumption we ensure that a correct mass-balance is always obtained. This is particularly interesting for SCE as their goal is to create a consistent plan across the entire supply chain.

The modeller decides which building blocks are activated at a specific location for a stream in a certain time period. With the help of these building blocks the modeller

is able to set up the network that represents the supply chain. Some examples of these building blocks are:

- Supply—Used to model that a certain stream is produced (becomes available) at a location in a given time period. Example: Purchase of a feedstock on the market;
- Storage—A stream can be put on stock (used) to use in a next time period or a stream can be taken from stock (become available) from a past time period. Example: Storage of a product on a production site;
- Transport—A stream can be transported into a location (become available) or transported out of a location (used). Example: Transport between locations that takes more than one month;
- Production—A stream can be produced during a process (become available) or it can be consumed during a process (used). These processes normally have fixed yields, i.e. the ratio in which streams are consumed and streams are produced is fixed. However, for some processes the yields depend on the utilisation of a production unit, which makes the model non-linear. Example: The production of the various units on the various sites;
- Interchangeability—Some stream can be interchanged (used) to another stream (become available). Example: More streams can be used for the same production process (for example a fuel), so they are interchangeable;
- Demand—Some streams (e.g. end-products) can be used to fulfil customer demand (used). Example: Demand for a product in certain demand areas.

At SCE the supply chain has been modelled with the above building blocks by several SME(s). The building blocks are activated by specifying certain parameters (usually a minimum a maximum or a cost/price). When modelling the supply chain the main challenge was to find the right level of detail such that it sufficiently reflects reality to provide answers, but also remains manageable (and solvable).

3.2 Operational Constraints

After setting up the supply chain various operational limits needed to be modeled. An example of such a limit at SCE is the capacity of a production unit. There are various ways to model these operational constraints in GMOS/NetSim. These constraints are activated by specifying the input parameters relevant to that particular constraint. Some basic examples of these constraints are:

- How much feedstock can we purchase? How much feedstock do we need to purchase contractually? (Building block: Supply)
- How much can we store in a tank? Is there a minimum level required? (Building block: Storage)
- How much can we transport to a location? How long does it take? Which types of transport are available? (Building block: Transport)

- What is the capacity of the production unit? Can it run on multiple modes? What will my yields be? (Building block: Production)
- How much fuel can be used to heat a production unit? (Building block: Interchangeability)
- How much product can we sell in a location next month? Is there a minimum/maximum? (Building block: Demand)

A lot of data is needed from several fields of expertise to model these constraints. Moreover, the input data needs to be updated on a regular basis as the supply chain changes over time. At SCE we succeeded to model the operational constraints for the relevant parts of the supply chain. Some parameters are updated on a yearly basis, where others are constantly monitored.

3.3 Economic Drivers

As the objective is to maximize the total integrated profit we have to assign costs and revenues to the various decisions that can be taken in the model. Examples are the cost for purchasing a feedstock, the revenue from selling a product or the cost to keep something in storage. Clearly, these economic drivers heavily influence the outcomes from the optimization. Moreover, the prices in the chemicals market constantly change. As such the prices have to be updated on a regular basis.

3.4 Detailed Margin Analysis

At SCE there was a need to be able to analyze the margins made on a feedstock or product on a more detailed level. In the main optimization we only focus on the total profit and cost, but not necessarily on the margin per product or feedstock.

Suppose for example that a production unit uses multiple feedstocks for which one feedstock may come from different sources (storage/market/internal production). Then we would like to assign part of the total costs of these feedstocks to each product produced in the production unit. These products might be used as a feedstock in another production process for which also other feedstocks with different costs are used. Again we would like to assign these costs in such a way that we know how much it has cost us to produce one ton of a certain product. In the same way we are interested in the total revenue that is eventually made on a certain feedstock after it has been converted to one or more products. This is done with a similar logic where revenues are assigned to the feedstocks. In such a way we are able to obtain a good insight in the margins made on products and feedstocks.

Please note that this analysis is done after the main optimization with a series of LP's. In the image below you can find an overview of the mathematical models that are solved in sequence each using the output of former model.

After solving the integrated profit optimization model (NLP) an LP is solved that helps us to determine how a feedstock (or product) flows through the network. This is the source for our margin calculation and helps in understanding the flows through the network. These flows through the network are determined on a more detailed level (within a location) and do not influence the results in the main optimizations.

Based upon these flows we solve a second LP that is used to assign the costs and revenues to the flows that have been determined. The key idea behind this LP is that we forward track the (cumulative) cost and that we backward track the revenue (minus cost) through the optimized network. Allocation of the costs is done with business rules for which several extensions have been created over the years.

The detailed margin analysis module in GMOS/NetSim is crucial for SCE as most of the decisions are made based on the margin of a feedstock or product. The breakdown of the total margin gives the SME an idea of where the actual profit is made and how we can enlarge that profit in future time periods. Moreover, due to the breakdown of margins decisions can be taken without re-running the model.

4 Conclusions

In order to be able to make optimal decisions across the entire value chain rather than local optimal decisions it was crucial to build a model for the entire supply chain. With a global model we are able to optimize the entire supply chain taking into account the same assumptions everywhere. We were able to model the supply chain with GMOS/NetSim which is a network optimization tool developed by Shell and ORTEC which is used in various businesses within Shell. A favorable feature of the tool is that it is relatively easy for a SME to set up a mathematical model without the need of extensive mathematical knowledge.

One of the key benefits of the model is that we are able to ensure a correct mass-balance and provide unified reports for the entire supply chain. Moreover, as margins are crucial for decisions-making at SCE a detailed margin analysis module has been built in GMOS/NetSim, which further improves decision-making.

One of the main challenges that we faced is to get all the input data together with sufficient quality. The input needs to be obtained from various people representing different products and locations. Here it is crucial that the input data is of sufficient

quality in order to take the right decisions. We were able to embed the model in the S&OP processes at SCE such that decisions are made based on fact based mathematical analysis. As such we believe that the SCE model is a great example of adding value by applying mathematical techniques in practice.

Time-Dependent Dynamic Location and Relocation of Ambulances

Lara Wiesche

Abstract The rescue service is an important part of public health care, which is provided to the general public by the state. A crucial aspect of the rescue service is the first aid of patients provided by the local emergency medical service (EMS). Given a limited budget, the available resources, e. g. ambulances, have to be used efficient in order to ensure a high quality coverage. Empirical studies have shown temporal and spatial variations of emergency demand as well as variations of travel times during the day. Existing models do not sufficiently consider time-dependency of important model parameters as demand and travel times for EMS vehicles. Especially the use of flexible ambulance locations, e.g. hospitals or voluntary fire departments, can be useful to reach a suitable coverage. A mixed-integer linear program is formulated in order to explicitly model time-dependent demand and travel times. On an extensive case study it is shown that the presented dynamic model outperforms existing static models with respect to coverage and utilization of resources.

1 Introduction

The rescue service is an important part of public health care, which is provided to the general public by the state. A crucial aspect of the rescue service is the first aid of patients provided by the local emergency medical service (EMS). The economic pressure of rising health care costs leads to cost-effective planning while a high quality of medical care for the population has to be ensured. A key challenge for emergency services planners is an efficient usage of the capacity that can be realized by a high utilization of the available resources [1]. Especially in Germany it is observable, that existing EMS systems are exposed to high economic pressure.

L. Wiesche (✉)
Faculty of Management and Economics, Chair of Operations Research and Accounting,
Ruhr University Bochum, 44780 Bochum, Germany
e-mail: lara.wiesche@rub.de

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_65,

Accordingly, it is necessary to modify existing systems without substantial budget increases.

The quality of medical care in the emergency service is measured by a response time interval: the time emergencies can be reached within a legal time frame. System performance is measured as the number (or fraction) of calls that can be reached within the fixed time frame, most commonly 90 % of calls in less than 9 min during a year [2]. Requirement for the achievement of patients within the time limit is the spatial and temporal availability of ambulances and qualified staff. The availability is mainly influenced by the emergency demand, the travel speed and the service time which cyclical vary during the day and influences the availability of ambulances. A mixed-integer linear program is formulated in order to explicitly model time-dependent demand and travel times. It is shown on large empirical data records that the presented dynamic model outperforms existing static models with respect to coverage and utilization of resources.

2 Time-Dependent Ambulance Location

On each planning level various approaches have been treated in the literature to improve the quality of emergency care [3]. As a substitute for the respond time threshold existing approaches in literature maximize the (double) coverage of emergency demand areas within the legal time frame [4]. Depending on the emergency demand and ambulance travel time an optimal allocation of ambulances must be guaranteed. Empirical studies show that demand changes spatial, mainly caused by the population density, as well as temporally, mainly caused by the activities of the people at this time. In addition, ambulance travel time variations, especially caused by varying the traffic volume, are observable which directly affects the availability of emergency vehicles. The existing quantitative models for tactical EMS resource planning are too restrictive in this respect, since dynamic influences are not considered sufficiently.

On the basis of the double standard model [5] and its extension [6] a new approach is presented which explicitly considers demand and travel time variations throughout the day. This enables a temporal differentiation of resources. Taking into account different requirements, the coverage of EMS demand areas is maximized. The binary decision variable x_{it}^{k} indicates, if demand node i is covered k times in time period t while y_{jt} represents the number of ambulances located at node j. The time-dependent travel time matrix is given through $\tau_{ijt} := \tau_{ij}(t) = d_{ij}/v_t$. Corresponding to the time-dependent travel time the set $\mathcal{N}_{it}^{k} := \{j \in \mathcal{J} \mid \tau_{ijt} \leq r_k,\ k \in \{1, 2\}\}$ indicates all vehicle locations from which a demand site i can be reached within a (time) radius r_k ($r_1 \leq r_2$) in period t, which represents the legal response time. The combination of constraints (1)–(2) ensures the necessary coverage that a proportion α of the total time-dependent demand (d_{it}) is covered within r_1 and the whole demand area is covered within r_2. Constraints (3)–(4) express that a demand node is only covered if there is an ambulance within the neighborhood N_{it}^{1} and can only be covered k-times if it is also covered $k-1$-times.

$$\sum_{j \in \mathcal{N}_{it}^2} y_{jt} \geq 1 \qquad \forall i \in \mathcal{I}, \forall t \in \mathcal{T} \tag{1}$$

$$\sum_{i \in \mathcal{I}} d_{it} x_{it}^1 \geq \alpha \sum_{i \in \mathcal{I}} d_{it} \qquad \forall t \in \mathcal{T} \tag{2}$$

$$\sum_{j \in \mathcal{N}_{it}^1} y_{jt} \geq x_{it}^1 + x_{it}^2 \qquad \forall i \in \mathcal{I}, \ \forall t \in \mathcal{T} \tag{3}$$

$$x_{it}^1 \geq x_{it}^2 \quad \forall i \in \mathcal{I}, \qquad \forall t \in \mathcal{T} \tag{4}$$

The dynamic allocation of ambulances is modeled in constraints (5)–(8). On the one hand dynamic allocation means relocation of ambulances during the day (u_{ijt}), but on the other hand it means a time dependent ambulance fleet size (p_t). Thereby the length of the time periods has to be chosen carefully, since they have to be small enough to aggregate an appropriate demand and travel-time average as well as large enough to avoid excessive relocations every hour.

Constraint (5) limits the total number of vehicles in period t to p_t, beside the maximizing of the coverage, provision costs are indirectly integrated in the new model. In order to model the different numbers of vehicles in the system according to the time period t a (fictive) depot node D is integrated. If an ambulance is placed at the depot it means, that the vehicle is not manned with staff. Constraint (6) gather accumulates the number of vehicles located at the depot in y_{Dt}. Equations (7) and (8) ensure resulting relocations of vehicles between different locations that can take place accordingly.

$$\sum_{j \in \mathcal{J}} y_{jt} \leq p_t \qquad \forall t \in \mathcal{T} \tag{5}$$

$$y_{Dt} = p_{ges} - p_t \qquad \forall t \in \mathcal{T} \tag{6}$$

$$y_{jt} + \sum_{i \in \mathcal{J} \cup \{D\}} u_{ijt} - \sum_{i \in \mathcal{J} \cup \{D\}} u_{jit} = y_{j(t+1)} \quad \forall j \in \mathcal{J} \cup \{D\}, \ \forall t \in \mathcal{T} \setminus \{T\} \tag{7}$$

$$y_{jT} + \sum_{i \in \mathcal{J} \cup \{D\}} u_{ijT} - \sum_{i \in \mathcal{J} \cup \{D\}} u_{jiT} = y_{j1} \quad \forall j \in \mathcal{J} \cup \{D\} \tag{8}$$

In contrast to the Anglo-American EMS-system in the Franco-German EMS-system ambulances are placed only at rescue stations [7] which are in general placed near the city center. To increase the flexibility of the EMS-system, additional flexible locations e.g. volunteer fire departments or hospitals are considered as potential ambulance locations. Constraints (9) and (10) ensure the maximum capacity of vehicles (θ_j) at location j and prohibits the location at flexible stations during the night.

$$y_{jt} \leq \theta_j \qquad \forall j \in \mathcal{J}, \forall t \in \mathcal{T} \tag{9}$$

$$y_{jT} = 0 \qquad \forall j \in \mathcal{F} \tag{10}$$

The double coverage is considered in the maximization function of the optimization model (11) as well as the minimization of the relocations and use of flexible ambulance locations. Integrating relocations and flexible ambulance locations lead to additional flexibility especially during rush hours. A trade-off between flexibility and practicability

Table 1 Dynamic optimization in relation to a static optimization, comparison of the coverage degree and relocations

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
Dynamic model						
Single coverage (%)	100	99.97	99.03	100	98.29	98.91
Double coverage (%)	100	99.97	98.95	100	98.21	98.64
Number of relocations	3	4	0	2	5	0
Static model						
Single coverage (%)	100	99.97	99.75	100	98.29	98.91
double coverage (%)	100	99.97	99.32	100	98.21	98.64
Number of relocations	4	6	6	4	5	0
Coverage improvement (dynamic vs. static)	0	0	+1.09 %	0	0	0
Relocation avoidance (dynamic vs. static)	+1	+2	+6	+2	0	0

of flexible sites is observable. The objective function guarantees that flexible locations and relocations are only used if the influence on the demand coverage is high enough. The decision maker can decide about the usage of flexible location and relocations directly in the model through penalty costs (β and γ).

$$\max \sum_{t \in \mathcal{T}} \left(\sum_{i \in \mathcal{I}} d_{it} x_{it}^2 - \beta \sum_{i \in \mathcal{J}} \sum_{j \in \mathcal{J}} u_{ijt} - \gamma \sum_{j \in \mathcal{F}} y_{jt} \right) \tag{11}$$

The proposed model supports local municipalities for locating ambulances on a tactical decision level. With the objective of maximizing the double coverage taking into account relocations and flexible locations, the presented model integrates the variation of the emergency demand and ambulance travel speed, dynamic adjustments and flexible ambulance locations. Through the simultaneous consideration of different dynamic aspects, the resources can be temporally differentiated what leads to an efficient use of resources.

3 Application and Results

In the evaluation of the proposed tactical ambulance location model an extensive case study was implemented. The advantages of the proposed model are shown in a real world case study form the city of Bochum (Germany) with more than 20.000 anonymous annual operations in form of a square grid ($1 \times 1\,\text{km}^2$) aggregated data set. The total planing horizon of 24 h is equally split into $T = 6$ time-periods with a length of 4 h. The division of the day allows to aggregate an appropriate average of demand and travel time as well as integrating staff scheduling. On the basis of various developed criteria for the evaluation of coverage models (including the empirically required coverage degree of each planning squares) it is demonstrated, that the presented dynamic optimization model with simultaneous maximization of the coverage and minimizing the number of replacements should be preferred to the current allocation of the ambulances (status quo).

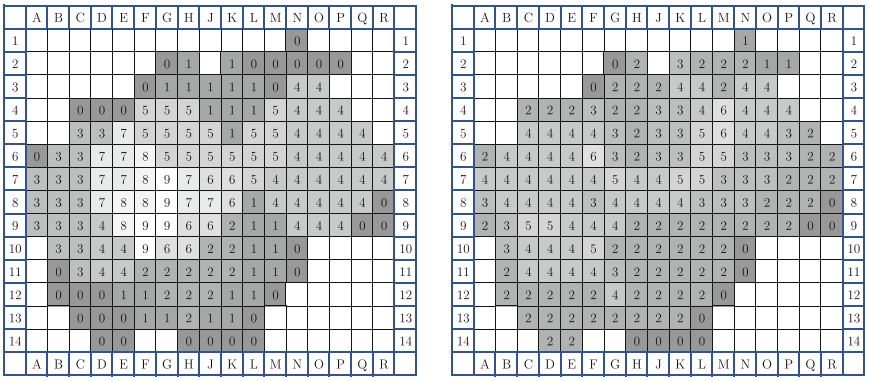


Fig. 1 Comparison of the coverage degree of each planning square in period 3 (8:00–12:00 a.m.), for the currently used allocation (*left*) and the optimal allocation (*right*)

A brief summary of the main results is provided in Fig. 1. The figure shows the coverage degree of each planning square in peak hours (8:00–12:00 a.m.), associated with high demands and low travel speed, for the currently used allocation of ambulances of the city of Bochum (left graph), compared to the optimal allocation of the model (right graph). From dark gray to light squares the number of coverage from zero to nine fold is visualized. Whereas dark gray squares indicate that emergency demand cannot be reached within the legal time frame (quality losses), the white squares represent an up to nine fold coverage which implies a waste of expendable resources. The comparison shows that the optimal allocation of ambulances in Bochum can achieve a much more uniform and thus more favorable coverage especially at peak hours. Among other factors the use of flexible locations in Bochum ensures an achievement of the outskirts during rush hours.

Additional flexibility of the solution is the permission of dynamic relocations during the day. In contrast to a static optimization model interdependences (placement of ambulances in an individual period will affect the placement in other periods) between the periods are considered in the dynamic model. The advantages of the dynamic model in comparison to a static model with $T = 6$ static period optimal solutions are shown in Table 1.

In contrast to the static (single) period model where the resulting number of relocations are not considered explicitly, the proposed dynamic model takes into account the cross-period relocations. As it can be seen the dynamic allocations correspond to the static one in terms of the solution quality—but at a significantly lower number of relocations. The analysis shows that the quality of the obtained solutions can be improved significantly by explicitly taking into account time-dependent variations in travel time, flexible ambulance locations and results in a better EMS supply.

4 Conclusion

With regard to the trend of increasing number of EMS operations, an efficient usage of existing emergency service resources is crucial. The simultaneous consideration of various dynamics and flexible ambulance locations, the resources can be temporally differentiated and used efficiently. The structure of the objective function allows the decision maker to directly influence the flexibility and therefore the solution of the EMS system. Especially the use of (existing) flexible location is an easy and inexpensive approach to improve the solution significantly.

References

1. Degel, D., Wiesche, L., Rachuba, S., & Werners, B. (2014). Reorganizing an existing volunteer fire station network in Germany. *Socio Economic Planning Sciences*, in press. doi:10.1016/j.seps.2014.03.001, http://www.sciencedirect.com/science/article/pii/S0038012114000147.
2. McLay, L. A., & Mayorga, M. E. (2010). Evaluating emergency medical service performance measures. *Health Care Management Science*, *13*(2), 124–136.
3. Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, *74*(3), 281–310.
4. Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research*, *147*(3), 451–463.
5. Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science*, *5*(2), 75–88.
6. Schmid, V., & Doerner, K. (2010). Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, *207*(3), 1293–1303.
7. Wolfgang, D. (2003). Anglo-American vs. Franco-German emergency medical services system. *Prehospital and Disaster Medicine*, *18*(1), 29–35.

Inventory Replenishment Models with Advance Demand Information for Agricultural Online Retailers

Haoxuan Xu, Yeming Gong, Chengbin Chu and Jinlong Zhang

Abstract This paper studies the inventory replenishment planning problems for agricultural online retailers able to obtain advance demand information (ADI) in an environment of time-varying demands. We incorporate ADI into dynamic lot-sizing (DLS) models to formulate the replenishment planning problems for agricultural online retailers. We consider three scenarios in this research. (1) Companies act as pure-play online retailers with customers homogeneous in demand lead time. (2) Online customers are heterogeneous in demand lead time with priorities. (3) Online retailers operate in a *bricks-and-clicks* structure, in which demands come from both online and offline channels. These channels can be either independent or interactive.

1 Introduction

This research is motivated by an inventory replenishment problem arising in an agricultural online retailer. Since most of its produce (e.g., fruits and vegetables) are perishable, this online retailer frequently procures them from a large agriculture products trade center located near its fulfillment center. The location privilege ensures timely and sufficient supply. In this environment of time-varying demands

H. Xu · J. Zhang
School of Management, Huazhong University of Science and Technology,
Wuhan 430074, China
e-mail: juwan.hsu@gmail.com

Y. Gong (✉)
EMLYON Business School, 69134 Ecully Cedex, France
e-mail: gong@em-lyon.com

C. Chu
Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes,
92295 Chatenay-Malabry Cedex, France

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_66,

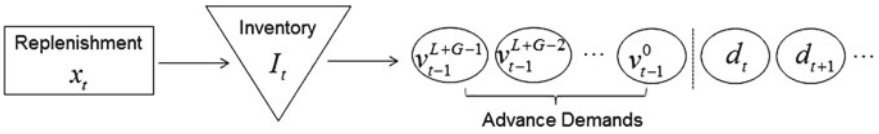


Fig. 1 Inventory model with ADI

for produce, we model the inventory replenishment plan of such agricultural online retailers as an uncapacitated single item lot sizing problem.

Agricultural online retailers obtain advance demand information (ADI) when customers place produce orders online, since these demands are usually be satisfied periods after the order place time. The time from a customer's order until the due date is defined as demand lead time (see [1]). We consider dynamic lot-sizing (DLS) models with ADI and flexible delivery under different online retailing scenarios to solve the practical problems. This study focuses on the inventory replenishment policy of the produce that can be easily ordered from suppliers or can be produced by the online retailers themselves.

2 Problem Formulation

2.1 DLS Model with ADI

In a rolling horizon environment, we consider an agricultural online retailer selling a kind of produce with timely and sufficient supply in the forthcoming periods from 1 to N. Demands arriving in each period $t\,(t = 1, \ldots, N)$ are supposed to be independent, and denoted as d_t. Given a standard demand lead time L, d_t will be satisfied within $[t, t + L]$ without any backorder penalty (see [4]). It can also be delayed by a maximum allowed time G, but with backlogging cost. Hence, the latest due date of d_t is $t + L + G$. We define $v_t^i (i = 0, \ldots, L + G - 1)$ as the unsatisfied part of the demand of i periods earlier at the end of period t. All such unsatisfied advance demands will be transferred to the next period, and the total quantity is I_t^-. Holding inventory at the end of period t is I_t^+.

In our model, d_t are known and deterministic, v_0^i are input data. Without loss of generality, we assume the holding inventory level at the beginning of the planning horizon is equal to zero, i.e. $I_0^+ = 0$. Figure 1 exhibits the inventory model with ADI. We first assume L to be homogeneous for all customers (see [2]). In period t, demands consist of two parts. One is the unsatisfied advance demands portfolio $\left(v_{t-1}^0, v_{t-1}^1, \ldots, v_{t-1}^{L+G-1}\right)$, the other part is d_t.

We also define the following:

k_t	fixed cost of ordering (set-up cost) in period t;
p_t	unit ordering/production cost in period t;
h_t	unit holding cost in period t;
b_t	unit backlogging cost in period t;
f_t	fixed delay delivery cost in period t;
x_t	amount replenished in period t;
$y_t = 1$	if $x_t > 0$, and 0 otherwise;
M	an arbitrarily large number;

Using this notation, we formulate the DLS problem with ADI as follows:

Model 1:

$$Min \sum_{t=1}^{N} (k_t y_t + p_t x_t + h_t I_t^+ + f_t v_t^L + \sum_{m=0}^{G-1} b_t v_t^{L+m}) \tag{1}$$

subject to:

$$(I_t^+ - I_t^-) = (I_{t-1}^+ - I_{t-1}^-) + x_t - d_t, \quad t = 1, \ldots, N \tag{2}$$

$$I_t^- = \sum_{i=0}^{L+G-1} v_t^i, \quad t = 0, \ldots, N \tag{3}$$

$$I_{t-1}^+ + x_t - v_{t-1}^{L+G-1} \geq 0, \quad t = 1, \ldots, N \tag{4}$$

$$\sum_{m=i}^{L+G-1} v_t^m \geq \sum_{m=i-1}^{L+G-1} v_{t-1}^m - x_t - I_{t-1}^+, \quad i = 1, \ldots, L+G-1 \tag{5}$$

$$\sum_{m=0}^{L+G-1} v_t^m \geq \sum_{m=0}^{L+G-1} v_{t-1}^m - x_t - I_{t-1}^+ + d_t \tag{6}$$

$$v_t^i \leq v_{t-1}^{i-1}, v_t^0 \leq d_t, \quad i = 1, \ldots, L+G-1; t = 1, \ldots, N \tag{7}$$

$$0 \leq x_t \leq M y_t \tag{8}$$

$$I_N^+ = I_N^- = 0 \tag{9}$$

$$v_t^i \geq 0, I_t^+ \geq 0, I_t^- \geq 0, \quad i = 0, \ldots, L+G-1; t = 0, \ldots, N \tag{10}$$

The objective function, Eq. (1), is to minimize the total cost. In our model, Once an item is delivered after L periods, there is a fixed cost f_t charged at the first period of delay. For each period t during the delaying time, a cost b_t is also charged. constraints (2–3) are inventory balance equations, and (4) guarantees the unsatisfied demands with maximum allowed delay in period t are satisfied. Constraints (5–7) denote how the unsatisfied advance demands in period $t-1$ transfer to be that in period t. Difference is that when computing v_t^0 (unsatisfied demand in the current period t), d_t should be considered. Constraint (8) ensures the sufficient replenishment

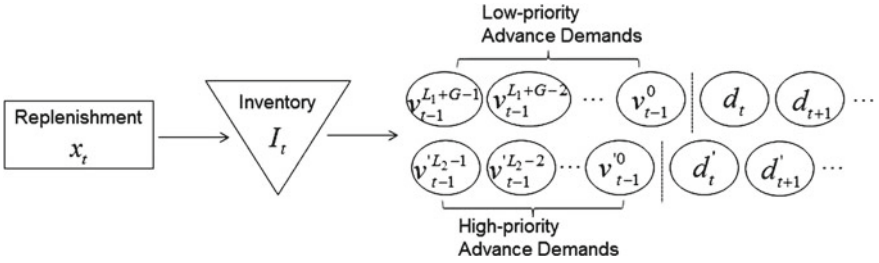


Fig. 2 Inventory model with online demand priority

quantity, and (9) guarantees all the demands be satisfied with no holding inventory in the end. Constraint (10) defines the variable types.

2.2 Demand Priority

Some online retailers distinguish demands into different priorities. According to delivery option, we assume orders to be high and low priority. The latter means demands with standard delivery option, while the former signifies those demands with urgent delivery option. The demand lead time of high priority demands (L_2) is shorter than that of low priority demands (L_1), i.e. $L_2 < L_1$. Besides, High priority orders can not be backlogged. Figure 2 exhibits the inventory model with demand priority. Demands of period t come from two classes of customers. Each can be divided into two parts, advance demands and demands arriving in period t.

Based on Model 1, we formulate the problem with demand priority as follows.

$v_t^i, v_t^{'j}$ at the end of period t, unsatisfied part of low priority demand of i periods earlier, unsatisfied part of high priority demand of j periods earlier;
$d_t, d_t^{'}$ demands arriving in t from low, high priority customers;

Model 2:

$$Min \sum_{t=1}^{N}(k_t y_t + p_t x_t + h_t I_t^+ + f_t v_t^{L_1} + \sum_{m=0}^{G-1} b_t v_t^{L_1+m}) \tag{11}$$

subject to:

$$(I_t^+ - I_t^-) = (I_{t-1}^+ - I_{t-1}^-) + x_t - d_t - d_t^{'} \tag{12}$$

$$I_t^- = \sum_{i=0}^{L_1+G-1} v_{t-1}^i + \sum_{j=0}^{L_2-1} v_{t-1}^{'j}, \quad t = 0, \ldots, N \tag{13}$$

$$I_{t-1}^+ + x_t - v_{t-1}^{L_1+G-1} - v_{t-1}^{'L_2-1} \geq 0 \tag{14}$$

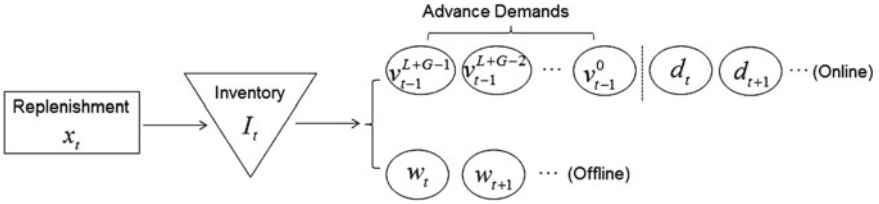


Fig. 3 Inventory model with independent demand channel

$$\sum_{m=i}^{L_1+G-1} v_t^m + \sum_{m=i}^{L_2-1} v_t^{'m} \geq \sum_{m=i-1}^{L_1+G-1} v_{t-1}^m + \sum_{m=i-1}^{L_2-1} v_{t-1}^{'m} - x_t - I_{t-1}^+,$$
$$i = 1, \ldots, L_2 - 1 \tag{15}$$

$$\sum_{m=i}^{L_1+G-1} v_t^m \geq \sum_{m=i-1}^{L_1+G-1} v_{t-1}^m + v_{t-1}^{'L_2-1} - x_t - I_{t-1}^+,$$
$$i = L_2, \ldots, L_1 + G - 1 \tag{16}$$

$$\sum_{m=0}^{L_1+G-1} v_t^m + \sum_{m=0}^{L_2-1} v_t^{'m} \geq \sum_{m=0}^{L_1+G-1} v_{t-1}^m + \sum_{m=0}^{L_2-1} v_{t-1}^{'m} - x_t - I_{t-1}^+ + d_t + d_t^{'} \tag{17}$$

$$v_t^i \leq v_{t-1}^{i-1}, v_t^0 \leq d_t, \quad i = 1, \ldots, L_1 + G - 1; t = 1, \ldots, N \tag{18}$$

$$v_t^{'j} \leq v_{t-1}^{'j-1}, v_t^{'0} \leq d_t^{'} \quad j = 1, \ldots, L_2 - 1; t = 1, \ldots, N \tag{19}$$

$$0 \leq x_t \leq M y_t \tag{20}$$

$$I_N^+ = I_N^- = 0 \tag{21}$$

$$v_t^i \geq 0, v_t^{'i} \geq 0, I_t^+ \geq 0, I_t^- \geq 0. \tag{22}$$

2.3 Demand Channel

2.3.1 Independent Channel

In a *bricks-and-clicks* online retailing structure, we first consider a scenario of independent demand channel. The aggregated demands in period t come from two independent channels, one is online store, the other is physical (offline) store. Figure 3 shows the inventory management process of independent demand channel. Demands in period t consist of advance demands from online customers and demands arriving from both offline and online customers. We use w_t to denote the offline demands, which must be satisfied immediately. Model 2 is applicable to this scenario when $L_2 = 0$.

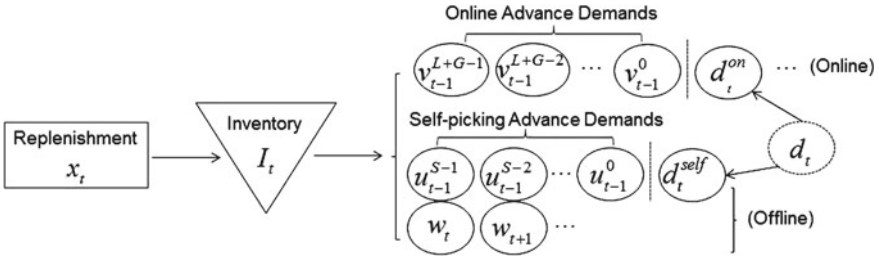


Fig. 4 Inventory model with interactive demand channel

Table 1 Numerical example for model 1

t	0	1	2	3	4	5	6	7	8	9	10	11	12
v_t^0	0	69	0	36	61	0	26	34	0	45	67	0	0
v_t^1	0	0	0	0	36	0	0	26	0	0	45	0	0
v_t^2	0	0	0	0	0	0	0	0	0	0	0	0	0
v_t^3	0	0	0	0	0	0	0	0	0	0	0	0	0
d_t	–	69	29	36	61	61	26	34	67	45	67	79	56
x_t	–	0	98	0	0	158	0	0	127	0	0	247	0
I_t^+	0	0	0	0	0	0	0	0	0	0	0	56	0
I_t^-	0	69	0	36	97	0	26	60	0	45	112	0	0

2.3.2 Interactive Channel

In the scenario of interactive demand channel, customers have two options for delivery after placing orders online. (a) to wait for the delivery home; (b) to go to the appointed physical store for the ordering products after being informed. Figure 4 shows such inventory management process.

We use d_t^{self} to denote the demands with self-picking, and d_t^{on} to denote demands of customers waiting for delivery. Different from advance demands online, the demands transferring from online to offline can not be backlogged, and must be satisfied within an allowed maximum waiting time S. We consider this scenario as a combination of the scenarios of demand priority and independent demand channel. u_t^j, $j = 0, \ldots, S-1$ (see Fig. 4) can be seen as $v_t^{'j}$ in model 2. By adding the demands coming from offline physical stores w_t to the left side of constraint (12), and the right side of constraints (14–17), we can use such modified model 2 to formulate the problem in scenario of interactive demand channel.

3 Solution and a Numerical Example

All the constraints in the models we formulate are linear. Hence, we use Cplex solver to solve these models by a mixed integer linear programming method. In addition, we can analyze the optimality properties to design polynomial algorithms to solve some large scale problems. In a numerical example, we use the data from [3]. Furthermore, we assume $L = G = 2$, $f_t = 1$, $b_t = 1$ in all periods, and the unsatisfied advance demands at the beginning $v_0^i = 0$. The result of model 1 is shown in Table 1.

Acknowledgments This research is supported by Collaborative Innovation Center for Modern Logistics and Business of Hubei (Cultivation), Modern Information Management Research Center (MIMRC) of HUST and NSFC (No.70901028; 71271095).

References

1. Hariharan, R., & Zipkin, P. (1995). Customer-order information, leadtimes, and inventories. *Management Science*, *41*(10), 1599–1607.
2. Huang, S., Axsäter, S., Dou, Y., & Chen, J. (2011). A real-time decision rule for an inventory system with committed service time and emergency orders. *European Journal of Operational Research*, *215*(1), 70–79.
3. Wagner, H. M., & Whitin, T. M. (1958). Dynamic version of the economic lot size model. *Management Science*, *5*(1), 89–96.
4. Wang, T., & Toktay, B. L. (2008). Inventory management with advance demand information and flexible delivery. *Management Science*, *54*(4), 716–732.

Coordinating a Three-Echelon Telecom Supply Chain with Spanning and Pair-Wise Revenue Sharing Contracts

Azarm Yeganehfallah, Hamid Mashreghi and Mohammad Reza Amin-Naseri

Abstract Nowadays, competitions between supply chains forces members to participate in strategic partnerships extending the isolated firm's competitive advantages. Contracting and in particular revenue sharing contract is one of the main applicable partnership mechanisms being dramatically analyzed in the literature for coordinating two-echelon supply chains. However there exist a handful of studies based on multi-echelon supply chains. Reviewing the literature, revenue sharing contracts can be developed through two approaches in multi-echelon supply chains: spanning and pair-wise schemes. In this research we review first, the last developments in telecom industries providing a new model for telecom supply chains, then we model different revenue sharing contracts in order to coordinate a three-echelon telecom supply chain facing demand uncertainty. Finally we compare the strengths and limitations for implementing different pair-wise and spanning revenue sharing contracts in telecom industries which can be helpful for both academics and practitioners.

A. Yeganehfallah · H. Mashreghi
CeTIM, LIACS, Leiden University, Leiden, The Netherlands
e-mail: a.yeganehfallah@umail.leidenuniv.nl; yeganehfallah@gmail.com

H. Mashreghi
e-mail: Mashreghi@Modares.ac.ir; h.mashreghi@umail.leidenuniv.nl

A. Yeganehfallah
Iran Company of Telecommunication, Babol, Iran

H. Mashreghi · M. R. Amin-Naseri (✉)
IE, TMU, Tehran, Iran
e-mail: Amin_nas@Modares.ac.ir

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_67,

1 Introduction and Background

Concerning growing logistic activities in the new economic world, the concept of coordination became interesting in practice as a paradigm to empower multi-echelon supply chain (SC) partnership for the chains want to be a market-leader. Reviewing the rich literature analyzing SC coordination, a handful of them were applied for multi-echelon SC with practical view. In this regard we aim to study telecom SC with its special features as a proved multi-echelon SC [1]. Supply chain coordination can be defined as finding the optimized decentralized decisions for SC partners aligned with optimized centralized SC. The main coordination mechanisms are classified by [2] namely; contracts, joint decision making, information sharing and information technology. Within this classification, contracts are one of the most relevant tools to achieve coordination from both academic and practitioners' point of view. Different types of contracts are analyzed in the literature for achieving coordination such as [3]: wholesale-price, returns-policy, revenue-sharing (RS), quantity discounts, quantity flexibility and sales rebate. Although the main focus of the literature is on analyzing two-echelon SC, we focus here on multi-echelon SCs.

1.1 Practical Relevancy in Telecommunication Industries

The rapidly changing atmosphere and landscape in telecommunication industry in over a hundred years of modern telecom history made a lot of shifts in the telecom supply chain. Telecommunication industry has a long history of focusing on reliability and performance but in closed network with new applications coming out at a snail's pace. On the other hand, internet is a young service providing an open environment producing a new model of communication full of innovative applications but with poor performance in real-time applications. However, with emerging Internet protocol multimedia subsystem (IMS) which merges internet world with telecom world, a new communications network will be produced offering a reliable network that will meet the needs of our real-time applications and at the same time provide an environment for new and innovative applications to meet the needs of end users. These new classes of services provides spaces for new entries [e.g. end-users service providers (SP)] in the telecom SC [11].

From the SC point of view, the classical telecom SC consists of a series of suppliers, an Electronics manufacturing service provider (EMS), an Original equipment manufacturer (OEM), and an operator as the last echelon which provides final products and services to end users [1]. However regarding the presence of IMS another final echelon can be considered between the classical last echelon, the operator, and the end-users (Fig. 1). This new player has a completely different nature compared to other ordinary telecom players. Firstly, this final echelon consists of multiple SP configuring in a horizontal shape vice versa to the SC meaning a user can get service from one to any number of them. Secondly, a variant of new services such as real-time

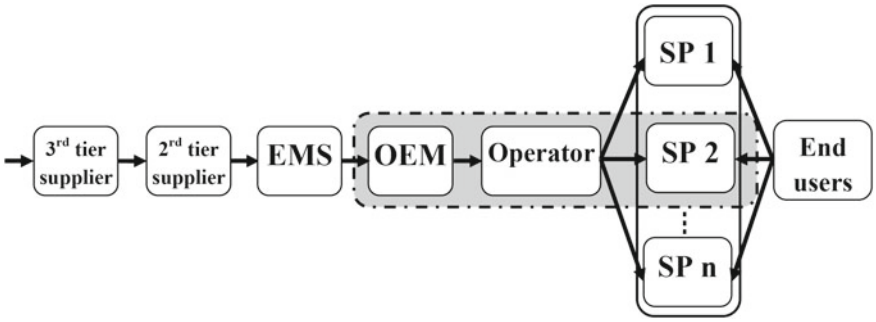


Fig. 1 Mapping a telecommunication multi-echelon SC: classical versus modern point of view

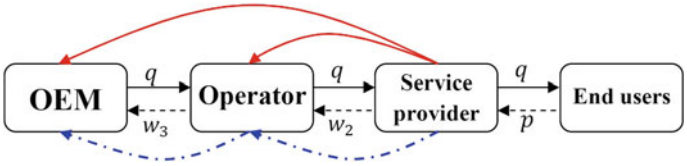


Fig. 2 Spanning (*upper arrows*) versus pair-wise (*dashed arrows*) RS contracts

health care, weather application combined with user location and news applications are available beside all renovated services of telecom and internet network. In this paper for simplification we focused only on one SP added to the total SC. Concerning the last three echelons of the telecom SC our model is based on the gray part of Fig. 1.

1.2 Literature Review: Multi-Echelon Supply Chain Contracting

Giannoccaro and Pontrandolfo [4] are the first ones who analyze RS contracts for three-echelon SCs in a Pair-wise setting with stochastic demand which is called here as PRS-1. Rhee et al. [9] argue that PRS-1 contract cannot coordinate SC and develop spanning RS (SRS) contract for a multi-echelon SC to achieve coordination (Fig. 2). Zhou and Yang [12] analyze three-echelon SC with a deterministic price-sensitive demand. However, in this paper we compare SRS contract with PRS-1 and a novel interpretation of pair-wise setting (PRS-2) under additive demand uncertainty. The other relevant research for multi-echelon SCs focus on wholesale-price contracts with tuning shortage cost [10], partial and complete information sharing [6], returns policies [5], and cooperative gaming with demand updates [7].

2 The Models and the Optimized Solutions

Assume a three-echelon telecom SC consisting of an OEM, an operator and a SP. This serial SC faces price-dependent additive uncertain demand, $D(p, x) = x + y(p)$, where $y(p) = a - bp$, and x is the random part with pdf, $f(x)$, and cdf, $F(x)$. There are many reasons for assuming demand uncertainty in telecom SC regarding emerging new operators, introducing new technologies, unpredictable growing demand, and appearance of new consumers geographically or in consumer segments [1]. For analysis of telecom services it can be assumed that any unmet demand is lost. Furthermore, no goodwill penalty cost is considered for shortages. Although risk analysis is an important issue in telecom industries, we assume SC partners are risk-neutral for simplification. Consider SP can order the quantity, q, before selling season starts. For simplicity we analyze SC coordination when p is fixed endogenously by market competition. Assume c_i: $1 \ldots 3$ as unit marginal costs for procurement of ordering quantity for SP, Operator, and OEM.

Defining stocking decision as $z = q - y(p)$, the SP's expected profit function (EPF) can be developed as $E(\Pi_1(z, p)) = p[y(p) + \mu] - (c_1 + w_2)[y(p) + z] - p\Theta(z)$ where $\Theta(z) = \int_z^B (z - x) f(x) dx$. It is interpreted as sum of riskless profit, the procurement costs regarding the level of q, and the transaction costs shortages occur. The EPF of the operator and the OEM become $E(\Pi_2(z, p)) = (w_2 - w_3 - c_2)[y(p) + z]$, and $E(\Pi_3(z, p)) = (w_3 - c_3)[y(p) + z]$. Therefore the SC's EPF becomes $E(\Pi_{SC}(z, p)) = p[y(p)+\mu]-c[y(p)+z]-p\Theta(z)$ where $c = \Sigma_{i:1\ldots3} c_i$. Considering a given p and strictly concavity of $E(\Pi_{SC}(z, p))$ for the most known density functions have non-decreasing hazard rates [8], the SC's first-optimality condition in z results in $F(z^*_{SC}) = \frac{(p-c)}{p}$. Similarly the SP's optimal z is $F(z^*_1) = \frac{(p-(c_1+w_2))}{p}$. To achieve coordination the sufficient condition $z^*_{SC} = z^*_1$ results in $c = c_1 + w_2$. Thus the wholesale-price agreement between SP and the operator cause $w_2 = c_2 + c_3$. It means the operator and the OEM obtain nonzero profits as a whole. However, it necessarily does not make the operator or the OEM having nonzero profit because it is possible to have an operator with positive profit which directs the entire negative profits to the OEM and vice versa. Nevertheless, this forces us to seek other coordinating schemes.

2.1 The Spanning Revenue Sharing Contract

With SRS contract (Fig. 2) it is assumed that the retailer assigns a part of selling revenue to the other partners [9]. Let ϕ^s_2, ϕ^s_3, and $1 - \phi^s_2 - \phi^s_3$ as the revenue shares for the operator, the OEM and the SP. SRS contract can redistribute the riskless profits through the operator and the OEM transferring the risk of losses for the case of shortages among all the partners. Let's rearrange the SC partners' EPFs:

$$E(\Pi_1^s(z,p)) = (1-\phi_2^s-\phi_3^s)p[y(p)+\mu] - (c_1+w_2^s)[y(p)+z] - (1-\phi_2^s-\phi_3^s)p\Theta(z),$$
$$E(\Pi_2^s(z,p)) = [\phi_2^s p[y(p)+\mu] - (c_2-w_2^s+w_2^s)[y(p)+z] - \phi_2^s p\Theta(z), \text{ and}$$
$$E(\Pi_3^s(z,p)) = [\phi_3^s p[y(p)+\mu] - (c_3-w_3^s)[y(p)+z] - \phi_3^s p\Theta(z).$$

Thus regarding the first optimality conditions the optimal ordering quantities become $F(z_{1,s}^*) = \frac{((1-\phi_2^s-\phi_3^s)p-(c_1+w_2^s))}{(1-\phi_2^s-\phi_3^s)p}$, $F(z_{2,s}^*) = \frac{(\phi_2^s p-(c_2-w_2^s+w_3^s))}{(\phi_2^s p)}$, and $F(z_{3,s}^*) = \frac{(\phi_3^s p-(c_3-w_3^s))}{(\phi_3^s p)}$.

2.2 The Pair-Wise Revenue Sharing Contracts

Regarding the pair-wise RS setting (Fig. 2) two different cases can be interpreted namely PRS-1 and PRS-2. Under PRS-1 which is introduced [4] and criticized [9] in the literature, the SP assigns ϕ_2^p share of the selling revenue to the operator and the operator similarly assigns ϕ_3^p shares of its total revenue to the OEM. The operator's total revenue includes the assigned selling revenue by the SP and also the earned revenue by selling services to the SP before the selling season. Thus the revenue shares of the SP, the operator and the OEM become respectively $1-\phi_2^p$, $(1-\phi_3^p)\phi_2^p$, and $\phi_3^p\phi_2^p$ and the EPFs can be rearranged as follows:

$$E(\Pi_1^p(z,p)) = (1-\phi_2^p)p[y(p)+\mu] - (c_1+w_2^p)[y(p)+z] - (1-\phi_2^p)p\Theta(z),$$
$$E(\Pi_2^p(z,p)) = (1-\phi_3^p)\phi_2^p p[y(p)+\mu] - (c_2-(1-\phi_3^p)w_2^p+w_3^p)[y(p)+z] - (1-\phi_3^p)\phi_2^p p\Theta(z), \text{ and}$$
$$E(\Pi_3^p(z,p)) = \phi_3^p\phi_2^p p[y(p)+\mu] - (c_3-w_3^p-\phi_3^p w_2^p)[y(p)+z] - \phi_3^p\phi_2^p p\Theta(z).$$

Accordingly the optimal ordering quantities become $F(z_{1,p}^*) = \frac{((1-\phi_2^p)p-(c_1+w_2^p))}{(1-\phi_2^p)p}$, $F(z_{2,p}^*) = \frac{((1-\phi_3^p)\phi_2^p p-(c_2-(1-\phi_3^p)w_2^p+w_3^p))}{((1-\phi_3^p)\phi_2^p p)}$, and $F(z_{3,p}^*) = \frac{(\phi_3^p\phi_2^p p-(c_3-w_3^p-\phi_3^p w_2^p))}{(\phi_3^p\phi_2^p p)}$.

Under another possible interpretation of pair-wise RS contracts which is named here PRS-2, the SP assigns ϕ_2^g share of the selling revenue to the operator. However, in this case the operator only assigns ϕ_3^g shares of its particular part of the SC's selling revenue to the OEM. Compared to PRS-1, the partners' revenue shares and the SP's EPF remain the same whereas the operators' and the OEM's EPFs are changed as follows:

$$E(\Pi_2^g(z,p)) = (1-\phi_3^g)\phi_2^g p[y(p)+\mu] - (c_2-w_2^g+w_3^g)[y(p)+z] - (1-\phi_3^g)\phi_2^g p\Theta(z), \text{ and}$$
$$E(\Pi_3^g(z,p)) = \phi_3^g\phi_2^g p[y(p)+\mu] - (c_3-w_3^g)[y(p)+z] - \phi_3^g\phi_2^g p\Theta(z).$$

Thus the optimal stocking decisions become $F(z^*_{1,g}) = \dfrac{((1-\phi_2^g)p-(c_1+w_2^g))}{(1-\phi_2^g)p}$, $F(z^*_{2,g}) = \dfrac{((1-\phi_3^g)\phi_2^g p-(c_2-w_2^g+w_3^g))}{((1-\phi_3^g)\phi_2^g p)}$, and $F(z^*_{3,g}) = \dfrac{(\phi_3^g\phi_2^g p-(c_3-w_3^g))}{(\phi_3^g\phi_2^g p)}$.

3 Analysis of Coordination

Toward finding the coordination conditions for SRS contract we should have $z^*_{SC} = z^*_{1,s}$, $z^*_{SC} = z^*_{2,s}$, and $z^*_{SC} = z^*_{3,s}$ which respectively results in $(\phi_2^s + \phi_3^s) = \frac{(c_2+c_3-w_2^s)}{c}$, $\phi_2^s = \frac{(c_2-w_2^s+w_3^s)}{c}$, and $\phi_3^s = \frac{(c_3-w_3^s)}{c}$. Thus it is feasible by setting appropriate revenue shares to achieve coordination by SRS contract. Similarly for PRS-1 contract we should investigate these conditions $\phi_2^p = \frac{(c_2+c_3-w_2^p)}{c}$, $(1-\phi_3^p)\phi_2^p = \frac{(c_2-w_2^p(1-\phi_3^p)+w_3^p)}{c}$, and $\phi_3^p\phi_2^p = \frac{(c_3-w_3^p-\phi_3^p w_2^p)}{c}$. From the first and the third conditions we have the sufficient condition $\phi_3^p = \frac{(c_3-w_3^p)}{(c_2+c_3)}$ satisfying all of them. Moreover, for PRS-2 contract the conditions $\phi_2^g = \frac{(c_2+c_3-w_2^g)}{c}$, $(1-\phi_3^g)\phi_2^g = \frac{(c_2-w_2^g+w_3^g)}{c}$, and $\phi_3^g\phi_2^g = \frac{(c_3-w_3^g)}{c}$ should be investigated. Accordingly all the conditions are satisfied with the sufficient condition $\phi_3^g = \frac{(c_3-w_3^g}{(c_2+c_3-w_2^g)}$.

4 Conclusion and Complementary Future Research

Coordination conditions shows for SRP and PRS-2 contracts, the SP and the OEM will be respectively the focal points for decision making, while PRS-1 contract fails coordination due to the separate optimization conditions of the partners. Therefore using PRS-2 contract not only resolves the barriers of PRS-1 contract but also provide a meaningful result for real SCs with dominant upstream partners. Based on the different services of a Telecom SC, the power of the partners would be changed over downstream to upstream. For instance, for high-tech services (e.g. IPTV, 3G and 4G mobiles,) the upstream is the dominant SC partner where for public services (e.g. healthcare, e-education, transportation,) the downstream i.e. SPs is the dominant partner. Moreover PRS-2 needs tuning the partners relationships peer to peer which is more convenient compared to SRS contract. Considering joint optimization of ordering and pricing, SRPs are more concerned because literature assumes that the retailer is responsible for setting selling price. Otherwise, if upstream partners can participate in pricing process (e.g. for high-tech services), PRS-2 contracts can be implemented to achieve coordination. Thus it is interesting to analyze the effect of pricing decisions on the ability of different RS contracts to coordinate a multi-echelon

SC. As another emerging issue, analyzing the competition between SPs should be analyzed simultaneously with coordination goals in telecom SCs.

References

1. Agrella, P. J., Lindroth, R., & Norrman, A. (2004). Risk, information and incentives in telecom supply chains. *International Journal of Production Economics*, *90*, 1–16.
2. Arshinder, K. A., & Deshmukh, S. G. (2008). Supply chain coordination: Perspectives, empirical studies and research directions. *International Journal of Production Economics*, *115*, 316–335.
3. Cachon, G. P. (2003). Supply chain coordination with contracts. In T. de Kok & S. Graves (Eds.), *Handbooks in operations and management science: Supply chain optimization*. The Netherlands: North-Holland Publishers.
4. Giannoccaro, I., & Pontrandolfo, P. (2004). Supply chain coordination by revenue sharing contracts. *International Journal of Production Economics*, *89*, 131–139.
5. He, Y., & Zhao, X. (2012). Coordination in multi-echelon supply chain under supply and demand uncertainty. *International Journal of Production Economics*, *139*, 106–115.
6. Jeong, I. J., & Leon, V. J. (2012). A serial supply chain of newsvendor problem with safety stocks under complete and partial information sharing. *International Journal of Production Economics*, *135*, 412–419.
7. Özen, U., Sošic, G., & Slikker, M. (2012). A collaborative decentralized distribution system with demand forecast updates. *European Journal of Operational Research*, *216*, 573–583.
8. Petruzzi, N., & Dada, M. (1999). Pricing and the newsvendor problem: A review with extensions. *Operations Research*, *47*, 183–194.
9. Van Der Rhee, B., Van Der Veen, J. A. A., Venugopal, V., & Nalla, V. R. (2010). A new revenue sharing mechanism for coordinating multi-echelon supply chains. *Operations Research Letters*, *38*, 296–301.
10. Seifert, R. W., Zequeira, R. I., & Liao, S. (2012). A three-echelon supply chain with price-only contracts and sub-supply chain coordination. *International Journal of Production Economics*, *138*, 345–353.
11. Wuthnow, M., Stafford, M., & Shih, J. (2010). *IMS a new model for blending applications* (1st ed.). New York: Auerbach Publications, Taylor and Francis Group.
12. Zhou, Y. W., & Yang, S. (2008). Pricing coordination in supply chains through revenue sharing contracts. *Information and Management Sciences*, *19*(1), 31–51.

The Material Loss and Failure Process in Sugar Production in Indonesia: A Case

Henry Yuliando, Adi Djoko Guritno and Endy Suwondo

Abstract The sugar produced in Indonesia is mainly from sugarcane. In its production, one of the problem is due to losses during the process. As found in this study, a material loss of sugar occured mainly in the milling plant. Based on data taken from a sugar plant in Yogyakarta, Indonesia, on May–June 2012, the left residue (bagasse) was 32 % of total input. The main cause was identified stem from the old machines and facilities that are used to experience a failure, and the lack of concern to the maintenance activity. As a solution, the company addressed a plan for maintenance that was divided into major and minor maintenance. Here, in this study, it propose a measurement on material loss, and see it correlation to the mean time between failure (MTBF) measurement. Since the milling plant is a continuous process, it is necessary to define a utilization capacity. This measure employees the availability schedule of the milling plant for every period (month) reduced by maintenance time that should be done in the middle of the process due to the encountered failure as estimated by MTBF. A correlation test was done to see whether there is a correlation between material loss and a verified utilization time. The result shows that there is a significant correlation, and discuss a need to put such problem in a platform of reliability management.

H. Yuliando (✉) · A. D. Guritno · E. Suwondo
Department of Agroindustrial Technology, Gadjah Mada University, Yogyakarta, Indonesia
e-mail: henry@tip-ugm.org

A. D. Guritno
e-mail: adidjoko@tip-ugm.org

E. Suwondo
e-mail: endys@gadjahmada.edu

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8_68,

1 Background

Sugarcane processing plant is one of the most important industries in Indonesia, mainly due to its role in the food safety for national consumption. Te industry tends to experience a decreasing productivity due to several factors, including the increasing plant age, a decline in the efficiency of plant equipment that requires replacement is constrained by the limited availability of investment capital. In general, the problems faced by the sugar industry covering issues on-farm and off-farm. Directorate General of Plantations (2010), states that on-farm problems are quite prominent, the sugarcane harvesting reached only around 6 tonnes/ha, and the sugarcane farming in the Java Island began to shift to other commodities. In the off-farm problem, sugar plant efficiency levels (overall recovery) which was still far below the standard, exacerbated by a relatively high production costs, and low levels of factory automation. The product quality is also relatively low, and yet the development of sugarcane-based product diversification if left behind than other producer countries. For instance, a sugar plant located Yogyakarta, PG Madukismo, as the only sugar factory in Yogyakarta, has been facing such problems. In efforts to meet the demand, PG Madukismo has been trying to increase sugar production by improving the performance of his sugar factory. One things that has to be done is by maintaining the difference between the material used (inputs) to output or commonly is referred to as material loss. Here, the term of material loss if known as loss of sugar caused during its processing.

Based on a preliminary observation, there were three major potential loss of sugar during the production process, i.e. in the producing of bagasse (by-product from milling process), filter cake (by-product from filtering), and mollase (by-product from centrifuge process). The sugar loss in those residue can be traced by the increasing in the value of the sugar content contained therein, stated with pol percentage value that each has a standard (bagasse has a standard of $\leq 2\,\%$). The complete sugar processing diagram in PG Madukismo can be seen in Fig. 2.

During the period of 2012 (May–September) the material loss or sugar loss in PG Madukismo had an increasing trend.

There is a very few literatures discussed about material loss in sugar processing. Cauchan et al. [2] conducted a research on life cycle assesment of sugar industry. His study resulted some input resources like power plant and distillery for optimal utilization of waste produced in sugar industry regarding to the environmental effects. Jorge et al. (2010) studied the evaporation of sugarcane juice in sugar production where the failure performance of the process decreased the output. While Somsen and Capelle (2002) conducted a research on yield analysis of food industry including sugarcane as the case. A certain cause of sugar loss is quality of input material (sugarcane).

As seen on Fig. 1, the increasing trend of material loss at PG Madukismo during the year 2012 indicated a decline in the efficiency of plant performance. Those issues should be handled properly since it can lead to the rising of production cost. The

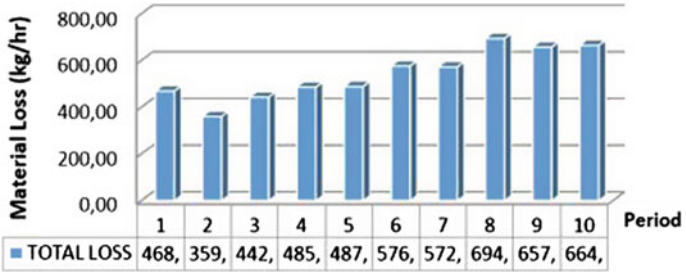


Fig. 1 Total loss in periods at PG Madukismo, Yogyakarta

inital step in proper handling is to find out the causes of the loss so as to know the steps in efforts to reduce material loss in sugar production.

As to the case of PG Madukismo, the material loss occurence is mainly caused by the utilization of the processing machines and facilities. We saw this was a potential cause due to the aging machine and facility used in the plant. In this matter the maintenance need to be managed accurately. As found in the preliminary survey, activities of maintenance in the company is divided into three types of maintenance, namely preventive, breakdown and corrective maintenance.

Those three types of maintenance are distinguished by the time it is carried out. The preventive maintenance is done prior processing activities, the breakdown maintenance is undertaken during the process. The failure of the process during operates can stop the process if it were fatal, and otherwise. Means that even it is said breakdown maintenance, it does not necessary to stop the process, but a maintenance is done while process continue. The sugar processing is type of continuous process. While corrective maintenance is aimed to give a priority over machine or facilities that according to the history data which had a high frequent to broken.

Here, as the purpose of this study, it is necessary for the company to determine mean time between failure (MTBF). As known that in term of MTBF, it is a prerequisite for the development of an effective preventive maintenance plan [1]. Smith [5], in his book stated that the MTBF indicates the availability time of the machine or facilities for processing with a constant rate of failure. And in this study the MTBF is analyzed to determine its effect on availability time for milling plant and drawing the correlation between the time of breakdown occured and the rate of material loss. Further, as seen on Fig. 2, the area of this study is mainly in the sugarcane milling process.

2 Materials and Method

In order to find the causes and measure the impact of the failure of the sugarcane milling process and the material loss, at first, the data collection of the scheduled preventive maintenance and the duration when the breakdown occured were recorded

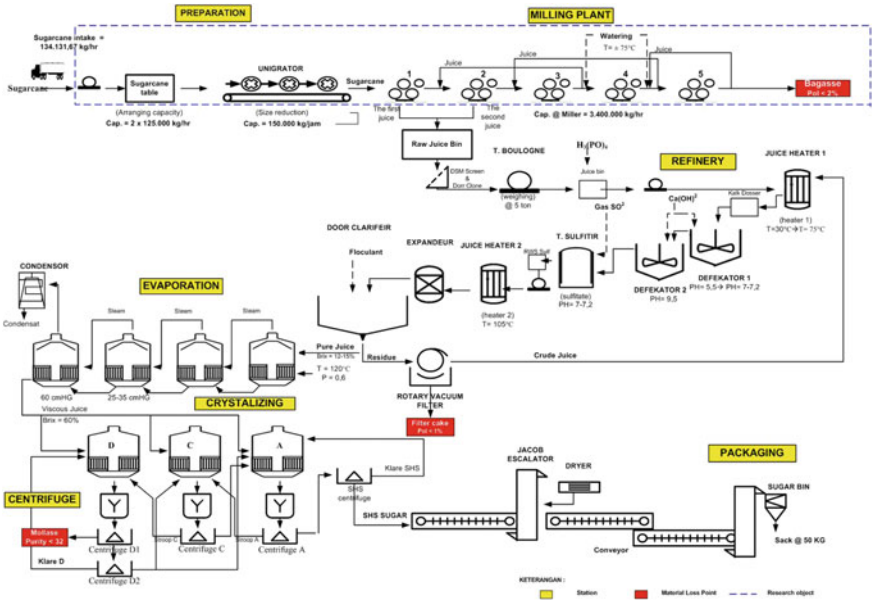


Fig. 2 The processing diagram of sugar plant at PG Madukismo, Yogyakarta, Indonesia

for periode May–September 2012. Over those data, MTBF is determined using a formula suggested by [3].

$$FR = \frac{number\ of\ failures}{\text{number of unit tested}} \times 100\ \% \tag{1}$$

$$FR(N) = \frac{number\ of\ failures}{\text{number of unit} - \text{hours of operating time}} \times 100\ \% \tag{2}$$

$$MTBF = \frac{1}{FR(N)'} \tag{3}$$

where $FR = product\ failure\ rate.$

Next, converting every unit to its availability time (operating) is calculated from total work days per period times 24 h. This amount then is reduced by the number of expected failure multiplied by maintenance duration time (on average) to determine the net operating time. The correlation is done between the rate of sugar loss against the real utilization (hours) of the milling plant. This measurements is used to draw a discussion in determining the causes of material loss, complemented by pareto diagram. And in-depth interview is conducted to enrich the analysis.

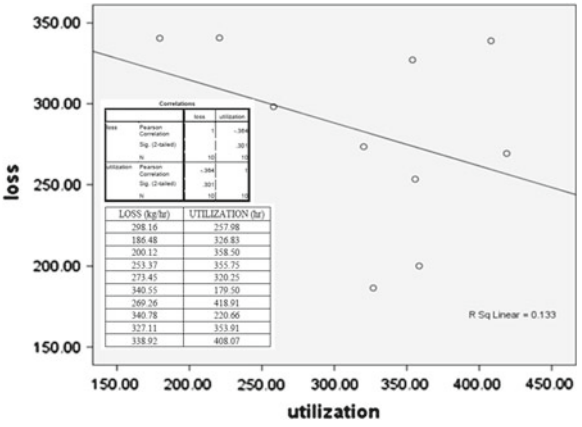


Fig. 3 The correlation between loss versus utilization

3 Result and Discussion

Based on the correlation test (see Fig. 3), the coefficient between los and utilizatioin is equal to −0.364 with a significance value of 0.301 (>0.05). It shows that the occurrence of material loss and utilization in the milling station (process) considerable influence each other in un-directional way, but not significant. This means that both variables affect each other quite negatively, i.e. if there is an increase in the utilization of the milling station, it will influence at 70 % degree to reduce the amount of material loss that occurs though statistically insignificant. This indicates other causes play the rest.

Based on the in-depth interview, the other factors that influence the occurrence of material loss are material input, worker, and method. Material input (sugarcane) consumed by the company has a wide variety that have a different properties. This is because PG. Madukismo taking sugarcane from different areas, particularly from surrounding regencies. The differences in variety and property of this material require different treatment and handling. However, the problem is particularly on the milling plant that cannot be adjusted to separate those material according to its properties (stiffness, tenacity, etc.) Due to the conitnuous process of the factory, the sugar loss varied in comply with the quality of the material.

Workers are also strongly assume affecting the rate of loss. Particularly in the view of different skill and educational background. This factors affect the worker performance. However since the process only run half a year, the company tend to have temporary worker rather permanently. The effort is by providing in-house for workers of fabrication and installation of each station before the milling season begin.

The working method, in contrast with a good manufacturing process, has been implemented in term of Standard Operating Procedure (SOP) that routinely updated

for by the company. However, sometimes in the presence unexpected situation, so that the SOP merely a formality because it is not able to resolve the existing problems.

Combining the above factors with the finding on MTBF in term of the process failure as implied in the duration (hours) for maintenance/repairing, all causes of material loss in sugar processing can be revealed. Here, achieving reliability, safety and maintainability results from activities in sugar processing are what the company should do. The reliability problem facing here can be categorized as realibility-focused operations [5]. It is explained by the situation of the plant that often stops and/or operates incorrectly, or beyond its operating limits, causes a higher failure rate. A reliability-focused operations team is necessary to enforce well-conceived standard operating procedures. The reliability-focused operations organization works closely with the maintenance team, particularly to provide inspection and operating health feedback on a regular basis. While not always maintenance can't improve the reliability of equipment, an inherent reliability should be recognized based upon design and operating context. The company can employ modern techniques like Reliability-Centered Maintenance (RCM), condition-based maintenance (CBM) and precision maintenance techniques. The organization works hard to optimize maintenance activities, with a focus on running time activities. It also works closely with operations to ensure that the equipment is available to produce as much product as required, and meet quality goals.

4 Conclusion

Material or sugar loss in sugar processing plant strongly rely on the reliability of the equipment used. The availability time of equipment should be adjusted by probability of failure. This adjustment can be constructed using the method of MBPF. In various way, the material input, worker and method also indicating an influence to the loss. The effort on managing reliability of the process based on operations and maintenance activities is necessary to overcome the loss, and can be supported by the program of RCM and/or CBM.

References

1. Braglia, M., Carmignani, G., Frosolini, M., & Zammori, F. (2012). Data classification and MTBF prediction with a multivariate analysis approach. *Reliability Engineering and System Safety*, *97*, 27–35.
2. Chauhan, M. K., Varun, S., Chaudhary, S., & Kumar, S. (2011). Life cycle assessment of sugar industry: A review. *Renewable and Sustainable Energy Reviews*, *15*, 3445–3453.
3. Directorate General of Plantations. (2010). *Road map blue print of national sugar self-sufficiency program*. Jakarta: Ministry of Agriculture.
4. Jorge, L.M.M., Righeto, A.R., Polli, P.A., Santos, O.A.A., & Maciel F, R. (2010). Simulation and analysis of a sugarcane juice evaporation system. *Journal of Food Engineering*, *99*, 351–359.

5. Smith, D. J. (2011). *Reliability, maintainability and risk: Practical methods for engineers* (8th ed.). UK: Butterworth-Heinemann.
6. Somsen, D., & Capelle, A. (2002). Introduction to production yield analysis - a new tool for improvement of raw material yield. *Trends in Food Science & Technology*, *13*, 136–145.

Author Index

D. Huisman et al. (eds.), *Operations Research Proceedings 2013*,
Operations Research Proceedings, DOI: 10.1007/978-3-319-07001-8,