

Karl-Heinz Waldmann
Ulrike M. Stocker
Editors

Operations Research Proceedings 2006



Gesellschaft für Operations Research e.V.



Österreichische Gesellschaft
für Operations Research



Schweizerische Vereinigung
für Operations Research



Springer

Operations Research Proceedings 2006

Selected Papers
of the Annual International Conference
of the German Operations Research Society (GOR),
Jointly Organized with the Austrian Society
of Operations Research (ÖGOR)
and the Swiss Society of Operations Research (SVOR)

Karlsruhe, September 6-8, 2006

Karl-Heinz Waldmann · Ulrike M. Stocker
(Editors)

Operations Research Proceedings 2006

Selected Papers
of the Annual International Conference
of the German Operations Research Society (GOR),
Jointly Organized with the Austrian Society
of Operations Research (ÖGOR)
and the Swiss Society
of Operations Research (SVOR)

Karlsruhe, September 6-8, 2006

With 123 Figures and 79 Tables

Prof. Dr. Karl-Heinz Waldmann

Universität Karlsruhe (TH)

Institut für Wirtschaftstheorie und Operations Research

Kaiserstraße 12

76131 Karlsruhe

Germany

e-mail: waldmann@wior.uni-karlsruhe.de

Dipl.-Wi.-Ing. Ulrike M. Stocker

Universität Karlsruhe (TH)

Institut für Wirtschaftstheorie und Operations Research

Kaiserstraße 12

76131 Karlsruhe

Germany

e-mail: stocker@wior.uni-karlsruhe.de

Library of Congress Control Number: 2007925442

ISSN 0721-5924

ISBN 978-3-540-69994-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Production: LE-TeX Jelonek, Schmidt & Vöckler GbR, Leipzig

Cover-design: WMX Design GmbH, Heidelberg

SPIN 11981145

42/3100YL - 5 4 3 2 1 0

Printed on acid-free paper

Preface

This volume contains a selection of papers referring to lectures presented at the symposium "Operations Research 2006" (OR 2006) held at the university of Karlsruhe, September 6 - 8, 2006. This international conference took place under the auspices of the Operations Research Societies of Germany (GOR), Austria (ÖGOR), and Switzerland (SVOR).

The symposium was attended by more than 600 academics and practitioners from 35 countries. It presented the state of the art in Operations Research and related areas in Economics, Mathematics, and Computer Science and demonstrated the broad applicability of its core themes, placing particular emphasis on Basel II, one of the most topical challenges of Operations Research.

The scientific program consisted of two plenary talks, eleven semi-plenary talks and more than 400 contributed papers, selected by the program committee and arranged in 19 sections. These presentations were complemented by the lectures of the GOR prize winners including the Unternehmenspreis, which has been awarded for the first time.

First of all we thank all participants of the conference, who submitted their paper for publication. However, due to a limited number of pages available for the proceedings volume, the total number of accepted papers had to be restricted.

Moreover, we want to express our thanks to the program committee and the section chairs for their support in acquiring interesting contributions and acting as a referee. Finally, we thank Anna Palej for gathering and editing the accepted papers as well as Dr. Werner A. Müller and Barbara Feß from Springer for their support in publishing this volume.

Committees

Program

I. Bomze, Universität Wien
H.W. Hamacher, Universität Kaiserslautern
T. Spengler, Technische Universität Braunschweig
H. Ulrich, Eidgenössische Technische Hochschule Zürich
K.-H. Waldmann, Universität Karlsruhe (TH), (chair)

Local Organization

K. Furmans, Institut für Fördertechnik und Logistiksysteme
W. Gaul, Institut für Entscheidungstheorie und Unternehmensforschung
A. Geyer-Schulz, Institut für Informationswirtschaft und -management
N. Henze, Institut für Stochastik
G. Last, Institut für Stochastik
D. Pallaschke, Institut für Statistik und Mathematische Wirtschaftstheorie
C. Puppe, Institut für Wirtschaftstheorie und Operations Research
O. Rentz, Institut für Industriebetriebslehre und Industrielle Produktion
K.-H. Waldmann, Institut für Wirtschaftstheorie und Operations Research
(chair)

Scientific Sections and Section Chairs

Business Intelligence, Forecasting and Marketing

Klaus Edel (Universität St. Gallen), Leena Suhl (Universität Paderborn)

Continuous Optimization

Christoph Helmberg (Technische Universität Chemnitz), Jean-Philippe Vial (Université Genève)

Discrete and Combinatorial Optimization

Thomas Liebling (ETH Lausanne), Uwe Zimmermann (Technische Universität Braunschweig)

Econometrics, Game Theory and Mathematical Economics

Günter Bamberg (Universität Augsburg), Ulrike Leopold-Wildburger (Universität Graz)

Energy and Environment

Karl Frauendorfer (Universität St. Gallen), Hans-Jakob Lüthi (ETH Zürich)

Finance, Banking and Insurance

Georg Pflug (Universität Wien), Hato Schmeiser (Universität St. Gallen)

Health and Life Science

Stefan Pickl (Universität Bw München), Marion S. Rauner (Universität Wien)

Logistics and Transport

Dirk Mattfeld (Technische Universität Braunschweig), Herbert Meyr (Wirtschaftsuniversität Wien)

Managerial Accounting and Auditing

Hans-Ulrich Küpper (Universität München), Thomas Pfeiffer (Universität Wien)

X Scientific Sections and Section Chairs

Metaheuristics and Decision Support Systems

Walter Gutjahr (Universität Wien), Stefan Voß (Universität Hamburg)

Multi Criteria Decision Theory

Matthias Ehrgott (University of Auckland), Christiane Tammer (Universität Halle-Wittenberg)

Network Optimization, Graphs and Traffic

Bettina Klinz (Technische Universität Graz), Anita Schöbel (Universität Göttingen)

Operational and Credit Risk

Thilo Liebig (Deutsche Bundesbank, Frankfurt a.M.), Wolfgang Stummer (Universität Erlangen-Nürnberg)

Production and Supply Chain Management

Hans-Otto Günther (Technische Universität Berlin), Richard Hartl (Universität Wien)

Revenue Management

Alf Kimms (Technische Universität Freiberg)

Scheduling and Project Management

Peter Brucker (Universität Osnabrück), Rainer Kolisch (Technische Universität München)

Simulation and Applied Probability

Nicole Bäuerle (Universität Karlsruhe), Ulrich Rieder (Universität Ulm)

Stochastic Programming

Petra Huhn (Technische Universität Clausthal)

System Dynamics and Dynamic Modelling

Gernot Tragler (Technische Universität Wien), Erich Zahn (Universität Stuttgart)

Contents

Part I GOR Unternehmenspreis 2006

Staff and Resource Scheduling at Airports

Ulrich Dorndorf 3

Part II GOR Dissertationspreis 2006

Produktionsplanung bei Variantenfließfertigung

Nils Boysen 11

Scheduling Buses and School Starting Times

Armin Fügenschuh 17

Dynamisches Bestandsmanagement in der Kreislauflogistik

Rainer Kleber 23

Periodic Timetable Optimization in Public Transport

Christian Liebchen 29

Determining SMB Superstructures by Mixed-Integer Optimal Control

Sebastian Sager, Moritz Diehl, Gundeep Singh, Achim Küpper, Sebastian Engell 37

Part III GOR Diplomarbeitspreis 2006

Complexity of Pure-Strategy Nash Equilibria in Non-Cooperative Games

Juliane Dunkel 45

Traffic Optimization Under Route Constraints with Lagrangian Relaxation and Cutting Plane Methods	
<i>Felix G. König</i>	53

Fare Planning for Public Transport	
<i>Marika Neumann</i>	61

Part IV Plenary and Semi-Plenary Talks

Recent Advances in Robust Optimization	
<i>Aharon Ben-Tal</i>	69

Neuro-Dynamic Programming: An Overview and Recent Results	
<i>Dimitri P. Bertsekas</i>	71

Basel II – Achievements and Challenges	
<i>Klaus Duellmann</i>	73

How to Model Operational Risk If You Must	
<i>Paul Embrechts</i>	81

Integer Quadratic Programming Models in Computational Biology	
<i>Harvey J. Greenberg</i>	83

On Value of Flexibility in Energy Risk Management. Concepts, Models, Solutions	
<i>Jörg Doege, Max Fehr, Juri Hinz, Hans-Jakob Lüthi, Martina Wilhelm</i> .	97

Bilevel Programming and Price Setting Problems	
<i>Martine Labbé</i>	109

Reliable Geometric Computing	
<i>Kurt Mehlhorn</i>	111

Financial Optimization	
<i>Teemu Pennanen</i>	113

Capital Budgeting: The Role of Cost Allocations	
<i>Ian Gow, Stefan Reichelstein</i>	115

An Overview on the Split Delivery Vehicle Routing Problem	
<i>Claudia Archetti, Maria Grazia Speranza</i>	123

Collaborative Planning - Concepts, Framework and Challenges	
<i>Hartmut Stadtler</i>	129

Promoting ϵ-Efficiency in Multiple Objective Programming: Theory, Methodology, and Application	
<i>Margaret Wiecek, Alexander Engau</i>	131

Part V Business Intelligence, Forecasting and Marketing

Combining Support Vector Machines for Credit Scoring	
<i>Ralf Steeking, Klaus B. Schebesch</i>	135

Nutzung von Data-Mining-Verfahren zur Indexprognose	
<i>Jonas Rommelspacher</i>	141

Zur Entscheidungsunterstützung bei netzeffektbasierten Gütern	
<i>Karl-Heinz Lüke, Klaus Ambrosi, Felix Hahne</i>	147

Part VI Discrete and Combinatorial Optimization

Nonserial Dynamic Programming and Tree Decomposition in Discrete Optimization	
<i>Oleg Shcherbina</i>	155

Mixed-Model Assembly Line Sequencing Using Real Options	
<i>Alireza Rahimi-Vahed, Masoud Rabbani, Reza Tavakkoli-Moghaddam, Fariborz Jolai, Neda Manavizadeh</i>	161

A New Approach for Mixed-Model Assembly Line Sequencing	
<i>Masoud Rabbani, Alireza Rahimi-Vahed, Babak Javadi, Reza Tavakkoli-Moghaddam</i>	169

On Asymptotically Optimal Algorithm for One Modification of Planar 3-dimensional Assignment Problem	
<i>Yury Glazkov</i>	175

A Multi-Objective Particle Swarm for a Mixed-Model Assembly Line Sequencing	
<i>Seyed Mohammed Mirghorbani, Masoud Rabbani, Reza Tavakkoli-Moghaddam, Alireza R. Rahimi-Vahed</i>	181

FLOPC++ An Algebraic Modeling Language Embedded in C++	
<i>Tim Helge Hultberg</i>	187

Two-Machine No-Wait Flow Shop Scheduling Problem with Precedence Constraints	
<i>Saied Samie, Behrooz Karimi</i>	191

A Multi-Commodity Flow Approach for the Design of the Last Mile in Real-World Fiber Optic Networks	
<i>Daniel Wagner, Günther R. Raidl, Ulrich Pferschy, Petra Mutzel, Peter Bachhiesl</i>	197

On the Cycle Polytope of a Directed Graph and Its Relaxations	
<i>Egon Balas, Rüdiger Stephan</i>	203

Modelling Some Robust Design Problems via Conic Optimization	
<i>Diah Chaerani, Cornelis Roos</i>	209

Polynomial Algorithms for Some Hard Problems of Finding Connected Spanning Subgraphs of Extreme Total Edge Weight	
<i>Alexey Baburin, Edward Gimadi</i>	215

Part VII Econometrics, Game Theory and Mathematical Economics

A Multidimensional Poverty Index	
<i>Gerhard Kocklaeuner</i>	223

Parameter Estimation for Stock Models with Non-Constant Volatility Using Markov Chain Monte Carlo Methods	
<i>Markus Hahn, Wolfgang Putschögl, Jörn Sass</i>	227

A Simulation Application for Predator-Prey Systems	
<i>Ulrike Leopold-Wildburger, Silja Meyer-Nieberg, Stefan Pickl, Jörg Schütze</i>	233

Robustness of Econometric Variable Selection Methods	
<i>Bernd Brandl</i>	239

Using Shadow Prices to Reveal Personal Preferences in a Two-Stage Assignment Problem	
<i>Anke Thede, Andreas Geyer-Schulz</i>	245

Part VIII Energy and Environment

Scheduling of Electrical Household Appliances with Price Signals	
<i>Anke Eßer, Andreas Kamper, Markus Franke, Dominik Möst, Otto Rentz</i>	253

Stochastic Optimization in Generation and Trading Planning

*Thomas Hartmann, Boris Blaesig, Gerd Hinüber,
Hans-Jürgen Haubrich* 259

Design of Electronic Waste Recycling System in China

Kejing Zhang, Daning Guo, Baoan Yang, Fugen Song 265

A Coherent Spot/Forward Price Model with Regime-Switching

Lea Bloechlinger 271

Part IX Finance, Banking and Insurance

A Management Rule of Thumb in Property-Liability Insurance

Martin Eling, Thomas Parnitzke, Hato Schmeiser 281

Heuristic Optimization of Reinsurance Programs and Implications for Reinsurance Buyers

Andreas Mitschele, Ingo Oesterreicher, Frank Schlottmann, Detlef Seese 287

Optimizing Credit Risk Mitigation Effects of Collaterals Under Basel II

Marc Gürtler, Dirk Heithecker, Martin Hibbeln 293

A New Methodology to Derive a Bank's Maturity Structure Using Accounting-Based Time Series Information

Oliver Entrop, Christoph Memmel, Marco Wilkens, Alexander Zeisler... 299

Sensitivity of Stock Returns to Changes in the Term Structure of Interest Rates – Evidence from the German Market

Marc-Gregor Czaja, Hendrik Scholz 305

The Valuation of Localization Investments with Real Options: A Case from Turkish Automotive Industry

Gül Gökay Emel, Pinar Özkeserli 311

Part X Health and Life Science

ILP Models for a Nurse Scheduling Problem

Bettina Klinz, Ulrich Pferschy, Joachim Schauer 319

Process Optimization and Efficient Personnel Employment in Hospitals

Gert Zülch, Patricia Stock, Jan Hrdina 325

Part XI Logistics and Transport

Inventory Control in Logistic and Production Networks

Bernd Scholz-Reiter, Salima Delhoum 333

Vehicle and Crew Scheduling with Flexible Timetable

András Kéri, Knut Haase 339

Lenk- und Ruhezeiten in der Tourenplanung

Asvin Goel, Volker Gruhn 343

Transport Channel Selection

Jürgen Branke, Denis Häußler, Christian Schmidt 349

A Sampling Procedure for Real-Life Rich Vehicle Routing Problems

Julia Rieck, Jürgen Zimmermann 355

Market-Oriented Airline Service Design

Cornelia Schön 361

‘T’ for Tabu and Time Dependent Travel Time

Johan W. Joubert 367

Integrated Operational Transportation Planning in Theory and Practice

Herbert Kopfer, Marta Anna Krajewska 373

Part XII Managerial Accounting and Auditing

Investment Incentives from Goal-Incongruent Performance Measures: Experimental Evidence

Markus C. Arnold, Robert M. Gillenkirch, Susanne A. Welker 381

Part XIII Metaheuristics and Decision Support Systems

Modelling Qualitative Information in a Management Simulation Game

Volker Nissen, Giorgi Ananidze 389

Schedule This - A Decision Support System for Movie Shoot Scheduling

Felix Bomsdorf, Ulrich Derigs, Olaf Jenal 395

A Framework for Truth Maintenance in Multi-Agent Systems <i>Brett Bojduj, Ben Weber, Dennis Taylor</i>	403

Part XIV Multi Criteria Decision Theory

Preference Sensitivity Analyses for Multi-Attribute Decision Support <i>Valentin Bertsch, Jutta Geldermann, Otto Rentz</i>	411
MCDA in Analyzing the Recycling Strategies in Malaysia <i>Santha Chenayah, Agamuthu Periathamby, Eiji Takeda</i>	417
Dimensionality Reduction in Multiobjective Optimization: The Minimum Objective Subset Problem <i>Dimo Brockhoff, Eckart Zitzler</i>	423
Multikriterielle Entscheidungsunterstützung zur Auswahl von Lagersystemen in der Ersatzteillogistik <i>Gerrit Reiniger, Martin Josef Geiger</i>	431

Part XV Network Optimization, Graphs and Traffic

Automatic Determination of Clusters <i>Bettina Hoser, Jan Schröder</i>	439
Online Dial-A-Ride Problem with Time Windows: An Exact Algorithm Using Status Vectors <i>Anke Fabri, Peter Recht</i>	445

Part XVI Operational and Credit Risk

Die Anwendung des Verlustverteilungsansatzes zur Quantifizierung operationeller Risiken <i>Frank Beekmann, Peter Stemper</i>	453

Part XVII Production and Supply Chain Management

Betriebskennlinien-Management als Performancemessungs- und -planungskonzept bei komplexen Produktionsprozessen <i>Dirk Eichhorn, Alexander Schömig</i>	461

Über verschiedene Ansätze zur Ermittlung von Betriebskennlinien – Eine Anwendungsstudie aus der Halbleiterindustrie <i>Alexander Schömig, Dirk Eichhorn, Georg Obermaier</i>	467
The Use of Chance Constrained Programming for Disassemble-to-Order Problems with Stochastic Yields <i>Ian M. Langella, Rainer Kleber</i>	473
Optimal Usage of Flexibility Instruments in Automotive Plants <i>Gazi Askar, Jürgen Zimmermann</i>	479
Comparison of Stochastic- and Guaranteed-Service Approaches to Safety Stock Optimization in Supply Chains <i>Steffen Klosterhalfen, Stefan Minner</i>	485
A Stochastic Lot-Sizing and Scheduling Model <i>Sven Grothklags, Ulf Lorenz, Jan Wesemann</i>	491
A Disassemble-to-Order Heuristic for Use with Constrained Disassembly Capacities <i>Tobias Schulz</i>	497
Supply Chain Management and Advanced Planning in the Process Industries <i>Norbert Trautmann, Cord-Ulrich Fündeling</i>	503
Production Planning in Dynamic and Seasonal Markets <i>Jutta Geldermann, Jens Ludwig, Martin Treitz, Otto Rentz</i>	509
A Branch and Bound Algorithm Based on DC Programming and DCA for Strategic Capacity Planning in Supply Chain Design for a New Market Opportunity <i>Nguyen Canh Nam, Le Thi Hoai An, Pham Dinh Tao</i>	515

Part XVIII Scheduling and Project Management

Branching Based on Home-Away-Pattern Sets <i>Dirk Briskorn, Andreas Drexel</i>	523
Priority-Rule Methods for Project Scheduling with Work Content Constraints <i>Cord-Ulrich Fündeling</i>	529

Entscheidungsunterstützung für die Projektportfolioplanung mit mehrfacher Zielsetzung	
<i>Antonia Maria Knübel, Natalia Kliewer</i>	535

Eine Web-Service basierte Architektur für ein Multi-Agenten System zur dezentralen Multi-Projekt Planung	
<i>Jörg Homberger, Raphael Vullriede, Jörn Horstmann, René Lanzl, Stephan Kistler, Thomas Göttlich</i>	541

Approaches to Solving RCPSP Using Relaxed Problem with Consumable Resources	
<i>Ivan A. Rykov</i>	547

Part XIX Simulation and Applied Probability

Risk-Sensitive Optimality Criteria in Markov Decision Processes	
<i>Karel Sladký</i>	555

Trading Regions Under Proportional Transaction Costs	
<i>Karl Kunisch, Jörn Sass</i>	563

Uniform Random Rational Number Generation	
<i>Thomas Morgenstern</i>	569

The Markov-Modulated Risk Model with Investment	
<i>Mirko Kötter, Nicole Bäuerle</i>	575

Optimal Portfolios Under Bounded Shortfall Risk and Partial Information	
<i>Ralf Wunderlich, Jörn Sass, Abdelali Gabih</i>	581

OR for Simulation and Its Optimization	
<i>Nico M. van Dijk, Erik van der Sluis</i>	587

Part XX Stochastic Programming

Multistage Stochastic Programming Problems; Stability and Approximation	
<i>Vlasta Kanková</i>	595

ALM Modeling for Dutch Pension Funds in an Era of Pension Reform	
<i>Willem K. Klein Haneveld, Matthijs H. Streutker, Maarten H. van der Vlerk</i>	601

Part XXI System Dynamics and Dynamic Modelling

**Identifying Fruitful Combinations Between System Dynamics
and Soft OR**
Myrjam Stotz, Andreas Größler 609

GOR Unternehmenspreis 2006

Staff and Resource Scheduling at Airports

Ulrich Dorndorf

INFORM — Institut für Operations Research und Management GmbH, Germany
ulrich.dorndorf@inform-ac.com

1 Introduction

At an airport, a large number of activities required for serving an aircraft while on the ground have to be scheduled. These activities include, for example, passenger and flight crew transportation, check-in and boarding services, various technical services, loading and unloading of cargo and baggage, or catering and cleaning services. With the steady increase of civil air traffic and the corresponding growth of airports over the past decades, the complexity of the task has increased significantly.

The scheduling process is mainly concerned with staff and terminal resources. Due to the very heterogeneous qualifications and properties required for performing the various services mentioned above, the scheduling problem is usually handled at the level of individual resource types such as, e.g., cargo and loading staff or aircraft stands. Schedules and corresponding staff and resource requirements are developed at the strategic, tactical and operational level.

The jobs or tasks to be scheduled depend on the flight schedule, which includes data on aircraft arrival and departure times, carrier, aircraft type, routing, passenger and freight numbers, etc., and on airline service agreements. The tasks have to be processed within certain time windows or fixed time intervals that follow from the aircraft arrival and departure times and from constraints within the aircraft turnaround process network. The tasks are obtained by applying a set of rules reflecting the service agreements to the flight schedule. They must be frequently updated on the day of operations: it is quite common that in a peak hour at a busy airport, several updates per second must be processed.

The following sections outline two models and solution approaches for task based staff shift scheduling and for scheduling terminal resources.

2 Task Based Shift Scheduling

Staff shift scheduling models are traditionally based on a demand curve representation of the required workload [5]. If task start and end times are fixed, this curve can be obtained by simply superimposing the tasks. Dantzig [2] has proposed the following model for the problem of covering the demand curve by a set of staff shifts with minimal cost (*demand based shift scheduling*):

$$\begin{aligned} \min \quad & \sum_{k \in K} c^k x^k \\ \text{s.th.} \quad & \sum_{k \in K} a_t^k x^k \geq d_t \quad \forall t \in T \\ & x^k \geq 0 \text{ and integer } \forall k \in K \end{aligned}$$

The integer decision variables x^k indicate the number of shifts of type k with an associated cost c^k . K denotes the set of shift types, T is the time horizon and d_t denotes the level of the labour demand curve for period t . A coefficient of the incidence matrix (a_t^k) takes the value 1 if period t is covered by shift type k and 0 otherwise.

Demand based shift scheduling works on the aggregated demand curve and ignores the aspect of assigning tasks to shifts. In many application areas, the demand curve is a natural representation of workloads, e.g. when the demand results from (random) customer arrivals and when the work force is homogeneous. In contrast to demand based shift scheduling, the following column generation model for *task based shift scheduling* simultaneously considers the allocation of tasks to shifts and the selection of an optimal number of shifts of each type:

$$\begin{aligned} \min \quad & \sum_{k \in K} c^k \sum_{p \in \Omega^k} \lambda_p^k \\ \text{s.th.} \quad & \sum_{k \in K} \sum_{p \in \Omega^k} a_{ip}^k \lambda_p^k = 1 \quad \forall i \in I \\ & \lambda_p^k \in \{0, 1\} \quad \forall k \in K, \forall p \in \Omega^k \end{aligned}$$

Here, the binary decision variables λ_p^k indicate whether a shift p of type k is part of the solution. An element of the incidence matrix (a_{ip}^k) takes the value 1 if shift p of type k covers work task i and 0 otherwise. I is the set of all work tasks and Ω^k denotes the index set for shifts of type k .

The task based shift scheduling model is similar to column generation models for vehicle routing, with a shift corresponding to a tour. In the model, an index $p \in \Omega^k$ describes a particular shift that contains a sequence of tasks. A shift is constructed in such a way that tasks do not overlap and are not preempted, and that it additionally contains relief breaks and respects other relevant constraints, e.g. sequence dependent setup times such as travel durations, or the qualifications required by the tasks within a shift. In an airport environment, these aspects make the model much more suitable for tactical and operational planning than the demand curved based model which does generally not directly lead to a workable plan.

The task based shift scheduling problem defined above is NP-hard in the strong sense. It can be solved through a branch and price approach [6]: starting with a set of heuristically generated shifts, additional shifts (columns) are generated in each iteration by solving a shortest path sub-problem; branching takes place over pairs of consecutive tasks that are covered on shifts with fractional shift variables. The approach can solve problem instances with more than 2000 tasks and up to 300 shift types in very moderate run times up to a few minutes on current PCs. Integral solutions are found quickly, and the LP gap is typically below 1%.

While the branch and price approach can solve many task-based scheduling problems occurring at airports to proven optimality within moderate run times, there are also important variants of the problem that call for a more complex objective function and require additional side constraints, e.g. for task start time windows, for absolute and/or relative minimum and maximum numbers of certain shift types used in a solution, incompatibilities of tasks within a shift, parallel placement of tasks and or shifts to allow for team work, buffer times between meal and relief breaks on a shift, etc.

Not all of these model extensions can be easily handled within the branch and price solution framework. We then reformulate the task based shift planning problem as a Constraint Programming model that is solved with a combination of truncated branch and bound and Large Neighbourhood Search (LNS), using the solution of the LP/IP relaxation for guiding the LNS.

3 Terminal Resource Scheduling

Terminal resource planning at an airport is concerned with the scheduling of immobile resources such as aircraft stands or gates, check in counters, waiting lounges, air-side terminal exits, baggage belts, etc. In contrast to staff and mobile equipment, the supply of these resources is usually fixed in the short and medium term, so that an optimal use becomes even more important. The scheduling of these resources is also particularly important from a passenger service point of view.

Models for terminal resource scheduling can be derived from the following basic fixed interval scheduling model:

$$\begin{aligned}
 & \max \sum_{i \in I} \sum_{j \in R} w_{ij} x_{ij} \\
 \text{s.th. } & \sum_{j \in R} x_{ij} \leq 1 \quad \forall i \in I \\
 & \sum_{i \in I(t)} x_{ij} \leq 1 \quad \forall j \in R, \forall t \in T \\
 & x_{ij} \in \{0, 1\} \quad \forall i \in I, \forall j \in R
 \end{aligned}$$

The binary decision variables x_{ij} take the value 1 if a task i is assigned to resource j and 0 otherwise; w_{ij} is the weight or preference for task to resource assignment. I denotes the set of tasks (that are in process at time t), R is the

set of resources, e.g. gates, and T is the time horizon. Fixed interval scheduling is NP-hard for general processing times or weights [1].

The IP model serves as a starting point for the development of specialised models. For example gate scheduling models may include special provision for the preemption of tasks to reflect the fact that aircraft may sometimes be towed between arrival and departure gates and add several specific constraints, e.g., for restricting the simultaneous assignment of large aircraft to neighbouring stands [3]. As another example, models for check in counter scheduling require that counters assigned to related tasks, e.g. for business and economy class check for the same flight, must be adjacent [4].

As terminal resource scheduling typically calls for multiple objective function criteria and requires several complicating side constraints, we have again found it helpful to reformulate the IP model as a Constraint Programming model. Using a solution approach based on truncated branch and bound and Large Neighbourhood Search, problem instances with up to 1000 tasks per day can be solved within very reasonable run times and with high solution quality (by comparison to current airport practices and to the off-line solution of simplified IP models).

4 Summary

Airport ground handling offers a great potential for optimisation at the level of strategic and tactical planning as well as in real time control. We have outlined two challenging problem classes for scheduling staff and terminal resources and have shown the underlying basic IP models. In our experience, the practical solution of these problems requires a combination of several modeling and solution techniques, including rule based systems for model and problem instance generation and Integer and Constraint Programming based solution techniques.

Given the frequency of changes in the underlying flight schedule, one of our major current developments consists in extending the models and algorithms in order to construct robust or stable schedules that take into account possible uncertainty or perturbations of the input data — being probably non-optimal in the original instance but as close as possible to the optimal one, optimizing for instance the worst case scenario.

References

1. Arkin EA, Silverberg EB (1987) Scheduling jobs with fixed start and end times. *Discrete Applied Mathematics* 18:1–8
2. Dantzig GB (1954) A comment on Edie's "Traffic delays at toll booths". *Journal of the Operations Research Society of America* 2:339–341
3. Dorndorf U, Drexel A, Nikulin Y, Pesch E (2007) Flight gate scheduling: State of the art and recent developments. *Omega* 35:326–334.

4. Duijn DW, van der Sluis E (2006) On the complexity of adjacent resource scheduling. *Journal of Scheduling* 9:49 – 62
5. Ernst AT, Jiang H, Krishnamoorthy M, Sier D (2004) Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research* 153:3–27
6. Herbers J (2005) Models and Algorithms for Ground Staff Scheduling on Airports. PhD Thesis, Rheinisch-Westfälische Technische Hochschule Aachen

GOR Dissertationspreis 2006

Produktionsplanung bei Variantenfließfertigung

Nils Boysen

Institut für Industrielles Management, Universität Hamburg, Germany
boysen@econ.uni-hamburg.de

1 Einleitung

Seit dem inzwischen schon legendär gewordenen Ausspruch von Henry Ford „Any customer can have a car painted any colour that he wants so long as it is black.“ hat ein fundamentaler Wandel bezüglich der Anforderungen an Produktionssysteme stattgefunden. So bietet heute etwa Daimler-Chrysler seine Mercedes C-Klasse aufgrund einer Vielzahl an vom Kunden individuell auswählbarer Optionen in 2^{27} theoretisch möglichen Varianten an [10]. Nichtsdestoweniger kann trotz dieser enormen Variantenvielfalt mittels Universalmaschinen mit automatisiertem Werkzeugwechsel und flexibel ausgebildeter Werker die effiziente Produktionsform der Fließfertigung aufrechterhalten werden. Eine solche Organisationsform der Fließfertigung, die eine Vielzahl an Varianten eines einheitlichen Grundmodells in wahlfreier Fertigungsfolge (Losgröße Eins) produzieren, bezeichnet man als Variantenfließfertigung. Man findet sie nicht nur bei der Endmontage von Autos und verwandten Produkten wie Bussen und sonstigen Nutzfahrzeugen, sondern auch in weiten Teilen der Elektroindustrie. Als Tribut an die gestiegene Variantenvielfalt muss jedoch eine größere Komplexität der Produktionsplanung in Kauf genommen werden. War es in den traditionellen Ein-Produkt-Fließsystemen mehr oder minder ausreichend eine einmalige Fließbandabstimmung bei der Installation des Fließsystems vorzunehmen, so treten bei einer Variantenfließfertigung gänzlich neue Planungsprobleme auf, deren hierarchisches Zusammenspiel in Abbildung 1 dargestellt ist [5]. Im Folgenden werden der Inhalt dieser einzelnen Planungsschritte beschrieben, einige aktuelle Forschungsergebnisse skizziert und wichtiger weiterer Forschungsbedarf benannt.

2 Fließbandabstimmung

Aufgabe der Fließbandabstimmung ist es, das Layout des Fließsystems zu bestimmen, dazu müssen die einzelnen Arbeitsgänge und die zur Durchführung

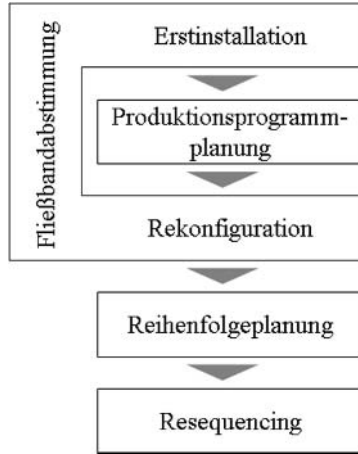


Fig. 1. Hierarchie der Produktionsplanung

benötigten Produktionsfaktoren den einzelnen Stationen zu geordnet werden. Dabei gilt es die Vorrangbeziehungen zwischen den Arbeitsgängen und eventuell weitere Nebenbedingungen wie etwa eine gegebene Taktzeit zu beachten. Die Fließbandabstimmung blickt auf eine über 50 jährige Tradition zurück. Trotzdem kann eine erhebliche Lücke zwischen den Ergebnissen der Forschung und den Bedürfnissen realer Fließsysteme konstatiert werden. Dies belegen einige ältere empirische Erhebungen aber auch ein jüngster Literaturüberblick [7], in dem aus der gesamten Literatur lediglich ein Anteil von unter 5% an Forschungsarbeiten identifiziert werden konnte, die sich explizit mit der Lösung von Praxisproblemen auseinandersetzen. Drei mögliche Gründe für diese Lücke und einige aktuelle Forschungsarbeiten zu deren Behebung seien an dieser Stelle genannt:

- Die Literatur hat unzählige praxisrelevante Erweiterungen der klassischen Fließbandabstimmung hervorgebracht. Zumeist werden diese Erweiterungen mit eigenen Namen versehen, die von Quelle zu Quelle variieren können. Um diesen „Wildwuchs“ zu strukturieren, findet sich in [6] eine Tupel-Klassifikation analog zur bekannten Scheduling-Klassifikation, mit deren Hilfe ein Großteil der Literatur zur Fließbandabstimmung mit dem jeweils behandelten Fließsystem erfasst wurde.
- Zumeist werden lediglich einzelne dieser Erweiterungen herausgegriffen und analysiert. Die Praxis benötigt jedoch viele dieser Erweiterungen in wechselnden Zusammensetzungen. Dementsprechend bedarf es flexibler Verfahren zur Fließbandabstimmung, die viele dieser Erweiterungen ohne größeren Anpassungsaufwand beachten können. Als eines der flexibelsten Verfahren hat sich in dem Literaturüberblick [6] das Verfahren Avalanche [1], [4] erwiesen.

- Schließlich erscheint es als ein erhebliches Problem, dass beinahe die gesamte Literatur (stillschweigend) von einer Erstinstallation eines Fließsystems ausgeht. Viel häufiger sind in der Praxis aber Rekonfigurationen von Fließsystemen durchzuführen. Die dabei etwa anfallenden Umzugskosten der Maschinen und Qualifizierungskosten der Werker sind bis dato noch nicht in der Literatur beachtet worden [7].

3 Produktionsprogrammplanung

In das Aufgabengebiet der Produktionsprogrammplanung fällt es, den vorliegenden Aufträgen einen Produktionstermin zunächst noch relativ grob, etwa auf Schichtebene zuzuordnen. Dabei gilt es sog. Abweichungskosten zu minimieren, die durch einen vom vereinbarten Liefertermin abweichenden Produktionstermin hervorgerufen werden. Zu beachten sind dabei etwa die Anzahl der Fertigungstakte einer Schicht, die Verfügbarkeit von Bauteilen aber auch die Belange der untergeordneten Reihenfolgeplanung. Planungsverfahren für diese Aufgabe sind bis dato kaum in der Literatur vorhanden. Erste Modelle und Algorithmen, die sich einer Umformulierung zum Multi-Resource Generalized Assignment Problem bedienen, finden sich in [1], [3] und [5].

4 Reihenfolgeplanung

Aufgabe der Reihenfolgeplanung ist es, den von der Produktionsprogrammplanung freigegebenen Aufträgen einer Schicht jeweils einen Fertigungstakt zuzuordnen. Dabei haben sich in der Literatur vor allem drei Klassen von Modellen etabliert [1], [2]:

- Das sog. Level-Scheduling entstammt dem Toyota-Production-System und zielt darauf, den durch die einzelnen Aufträge induzierten Materialbedarf möglichst gleichmäßig auf die Fertigungsfolge zu verteilen. Auf diese Weise soll dem Just-in-Time-Prinzip weitestgehend entsprochen und die Produktion in den einzelnen Fertigungsstufen ohne größere Lagermengen aufeinander abgestimmt werden. Ein aktueller Literaturüberblick findet sich in [8].
- Unterschiedliche Varianten benötigen in den einzelnen Stationen des Fließsystems aber auch unterschiedliche Bearbeitungszeiten. So macht es in der Automobilindustrie einen erheblichen zeitlichen Unterschied aus, ob ein elektrisches Schiebedach montiert wird, ein manuelles oder gar keines. Dementsprechend sollen aufwendige Varianten sich an allen Stationen möglichst mit weniger aufwendigen abwechseln, um eine Überlastung der Werker an den Stationen und damit Springereinsatz und/oder Qualitätsmängel zu vermeiden. Das sog. Mixed-Model-Sequencing trachtet danach, durch eine exakte zeitliche Terminierung der einzelnen Varianten an den Stationen diese Überlastungen exakt zu erfassen und zu minimieren.

- Schließlich will auch das sog. Car-Sequencing diese Überlastungen an den Stationen reduzieren, dabei jedoch den hohen Datenerhebungsaufwand des Mixed-Model-Sequencing umgehen. Dazu werden für einzelne Optionen Reihenfolgeregeln eingeführt. So besagt etwa eine Regel von 3:5, dass von fünf aufeinander folgenden Werkstücken maximal drei die entsprechende Option beinhalten dürfen. Das Car-Sequencing sucht nach Reihenfolgen, welche keine bzw. möglichst wenige der gegebenen Reihenfolgeregeln verletzt.

Speziell das Car-Sequencing hat in den letzten Jahren viel Beachtung in der Literatur gefunden. Dies lag unter anderem an der ROADEF-Challenge, bei der ein spezielles Car-Sequencing Problem des Automobilherstellers Renault zu lösen war. Bedienten sich frühere Veröffentlichungen vor allem dem Constraint-Programming zur Lösung des Car-Sequencing Problems, so haben neueren Veröffentlichungen im Zuge der ROADEF-Challenge vor allem leistungsfähige Meta-Heuristiken aber auch exakte Branch and Bound-Verfahren hervorgebracht [9].

5 Resequencing

Schließlich kann die geplante Reihenfolge häufig nicht in der Produktion eingehalten werden. Störungen wie Eilaufträge, Maschinenausfälle oder defekte Materialien behindern den Fertigungsablauf und führen zu "Verwirbelungen" der Fertigungsfolge. Eine gezielte Umstellung einer durch Störungen beeinflussten Fertigungsfolge im Laufe der Produktion bezeichnet man als Resequencing. Mit Hilfe von sog. Sortierspeichern kann etwa die Reihenfolge der Werkstücke physisch umgestellt werden, oder das Resequencing erfolgt lediglich virtuell, indem die Zuordnung zwischen Werkstück und Fertigungsauftrag getauscht wird [5]. Entscheidungsunterstützung für die Praxis findet sich in diesem Gebiet bis dato kaum.

6 Fazit

Abschließend lässt sich festhalten, dass jedes einzelne Planungsproblem noch viel Raum für weitere Forschungsarbeiten eröffnet. Da aber alle Einzelprobleme zahlreiche Interdependenzen untereinander aufweisen, sollten auch die Belange einer Hierarchischen Planung zukünftig verstärkt untersucht werden.

References

1. Boysen N (2005a) Variantenfließfertigung, DUV, Wiesbaden
2. Boysen N (2005b) Reihenfolgeplanung bei Variantenfließfertigung: Ein integrativer Ansatz. Zeitschrift für Betriebswirtschaft 75: 135–156

3. Boysen N (2005c) Produktionsprogrammplanung bei Variantenfließfertigung. Zeitschrift für Planung & Unternehmenssteuerung 16: 53–72
4. Boysen N, Fliedner M (2006) Ein flexibler zweistufiger Graphenalgorithmus zur Fließbandabstimmung mit praxisrelevanten Nebenbedingungen. Zeitschrift für Betriebswirtschaft 76: 55–78
5. Boysen N, Fliedner M, Scholl A (2006a) Produktionsplanung bei Variantenfließfertigung: Planungshierarchie und Hierarchische Planung. Jenaer Schriften zur Wirtschaftswissenschaft 22–2006
6. Boysen N, Fliedner M, Scholl A (2006b) A classification of assembly line balancing problems. Jenaer Schriften zur Wirtschaftswissenschaft 12–2006
7. Boysen N, Fliedner M, Scholl A (2006c) Assembly line balancing: Which model to use when? Jenaer Schriften zur Wirtschaftswissenschaft 23–2006
8. Boysen N, Fliedner M, Scholl A (2006d) Level-Scheduling: Klassifikation, Literaturüberblick und Modellkritik. Jenaer Schriften zur Wirtschaftswissenschaft 26–2006
9. Fliedner M, Boysen N (2006) Solving the Car Sequencing Problem via Branch & Bound. European Journal of Operational Research (to appear)
10. Röder A, Tibken B (2006) A methodology for modeling inter-company supply chains and for evaluating a method of integrated product and process documentation. European Journal of Operational Research 169: 1010–1029

Scheduling Buses and School Starting Times

Armin Fügenschuh

Fachgebiet Optimierung, Technische Universität Darmstadt, Germany
fuegenschuh@mathematik.tu-darmstadt.de

1 Introduction

Traffic peaks are peaks in cost. This in particular holds for rural counties, where the organization of public mass transportation is focused on the demand of pupils. About half to two third of pupils in rural areas take a bus to get to school. Most of them are integrated in the public bus system, a minority is transferred by special purpose school buses. In all cases the respective county in which the pupils live is responsible for the transfer, meaning that the county administration pays the fees. Since tax money is a scarce resource, the administration has great interest in reducing these payments.

A significant number of buses could be saved, if the bus scheduling problem is solved together with the starting time problem, i.e., the simultaneous settlement of school and trip starting times [6, 7]. A small intuitive example is shown in Figure 1. If two schools start at the same time then two different buses are necessary to bring the pupils to their respective schools. If they start at different times then one and the same bus can first bring pupils to one school and then pupils to the other. In this article we describe how to roll out this intuitive idea to a whole county. Besides presenting a mathematical formulation in Section 2 and computational results in Section 5, we want to

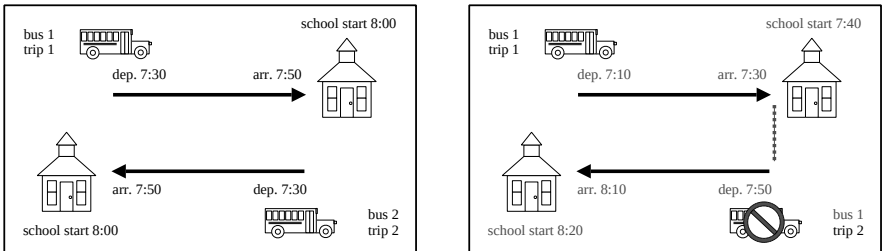


Fig. 1. The central idea (before – after)

put a particularly emphasis on the “legacy” of the PhD thesis [2]. That is, we want to point out those parts that are suitable for solving other combinatorial optimization problems: A new metaheuristic in Section 3 and a new preprocessing scheme in Section 4.

2 A Mathematical Model

Let $\mathcal{V}$ be the set of all passenger trips in the county under consideration. A *passenger trip* (or *trip* for short) $t \in \mathcal{V}$ is a sequence of bus stops, each having an arrival and a departure time assigned to. The time difference between the departure at the first and the arrival at the last bus stop is called the *trip duration*, and is denoted by δ_t^{trip} . (All time-related parameters and variables in this model are integral with the unit “minute”.) For every trip $t \in \mathcal{V}$ we introduce an integer variable $\alpha_t \in \mathbb{Z}_+$ representing its planned starting time, i.e., the departure of a bus at the first bus stop. A time window $\underline{\alpha}_t, \overline{\alpha}_t$ is given, in which the planned trip starting time must be, see inequalities (1f).

All buses start and end their tours at a *depot*. The trip to the first bus stop of trip t is called *pull-out trip*. When the bus arrives at the last bus stop of t , it is either sent on the *pull-in trip* back to the depot, or it serves another passenger trip. The duration of the pull-out and pull-in trip is denoted by $\delta_t^{\text{out}}, \delta_t^{\text{in}}$, respectively. The intermediate trip from the last bus stop of trip t_1 to the first bus stop of trip t_2 is called *deadhead trip*. The connection of a pull-out trip, several passenger and deadhead trips and a final pull-in trip which are then served by one and the same bus is called a *schedule*. For every trip $t \in \mathcal{V}$ the decision variables $v_t, w_t \in \{0,1\}$ indicate if trip t is the first or the last trip in some schedule, respectively. Let the set $\mathcal{A} \subset \mathcal{V} \times \mathcal{V}$ contain all pairs of trips (t_1, t_2) that can in principle be connected by a deadhead trip. The duration of the deadhead trip is given by $\delta_{t_1 t_2}^{\text{shift}}$. For every pair of trips $(t_1, t_2) \in \mathcal{A}$ the variable $x_{t_1 t_2} \in \{0,1\}$ indicates if t_1 and t_2 are in sequence in some schedule, that is, the same bus serves trip t_2 directly after finishing trip t_1 (apart from the deadhead trip and the idle time). Each trip is served by exactly one bus which is ensured by assignment constraints (1b). If trips $(t_1, t_2) \in \mathcal{A}$ are connected, then trip t_2 can only start after the bus has finished trip t_1 , shifted from the end of t_1 to the start of t_2 . Additional waiting is permitted if the bus arrives before the start of t_2 . Using a sufficiently big value for M , these constraints can be formulated as linear inequalities (1c).

Let $\mathcal{S}$ be the set of all schools in the county. For every school $s \in \mathcal{S}$ we introduce an integer variable $\tau_s \in \mathbb{Z}_+$. The school starting time is required to be in discrete time slots of 5 minutes (7:30, 7:35, 7:40, etc.). Thus the planned starting time of s is $5 \cdot \tau_s$. It is allowed to change this starting time within some time window $\underline{\tau}_s, \overline{\tau}_s$. Usually, the time window constraints (1f) reflect the legal bounds on the school starting time (7:30 to 8:30 a.m.).

The set $\mathcal{P} \subset \mathcal{S} \times \mathcal{V}$ consists of pairs (s, t) , where trip t transports pupils to a bus stop of school s . In this case we say t is a *school trip* for s . The time

difference between the departure at the first bus stop of t and the arrival at the bus stop of s is denoted by $\delta_{st}^{\text{school}}$. There is another time window for the pupils $\underline{\omega}_{st}^{\text{school}}, \bar{\omega}_{st}^{\text{school}}$, specifying the minimal and maximal waiting time relative to the school starting time. The lower bound $\underline{\omega}_{st}^{\text{school}}$ is chosen according to the walking time from the bus stop where the pupils leave the bus, whereas the upper bound $\bar{\omega}_{st}^{\text{school}}$ is due to law restrictions. For every $(s, t) \in \mathcal{P}$ the starting times of trip t and school s have to be chosen accordingly (1d).

Let $\mathcal{C} \subset \mathcal{V} \times \mathcal{V}$ be the set of pairs (t_1, t_2) , where t_1 is a trip that transports pupils to a so-called *changing bus stop*, where they leave the bus and transfer to trip t_2 . We say, t_1 is a *feeder trip* for t_2 and, vice versa, t_2 is a *collector trip* for t_1 . The driving time from the first bus stop of feeder trip t_1 to the changing bus stop is denoted by $\delta_{t_1 t_2}^{\text{feeder}}$. For the collector trip, the corresponding parameter is $\delta_{t_1 t_2}^{\text{collector}}$. At the changing bus stop, a time window $\underline{\omega}_{t_1 t_2}^{\text{change}}, \bar{\omega}_{t_1 t_2}^{\text{change}}$ for the minimal and maximal waiting time is given. For every $(t_1, t_2) \in \mathcal{C}$ the starting times of both bus trips must be synchronized in such way that t_2 arrives at the changing bus stop after trip t_1 within a small time window (1e).

The most important goal in the objective function (1a) is the reduction of the number of deployed vehicles (where C is the cost per bus), the second-most important is the reduction of the driving times of all pull-out, pull-in, and deadhead trips. For a discussion of multicriteria quality-of-service aspects we refer to [5]. Summing up, the model for the integrated optimization of school starting times and bus schedules looks as follows:

$$\min \sum_{t \in \mathcal{V}} C \cdot v_t + \sum_{t \in \mathcal{V}} \delta_t^{\text{out}} \cdot v_t + \sum_{(t_1, t_2) \in \mathcal{A}} \delta_{t_1 t_2}^{\text{shift}} \cdot x_{t_1 t_2} + \sum_{t \in \mathcal{V}} \delta_t^{\text{in}} \cdot w_t, \quad (1a)$$

$$\text{s.t.} \quad \sum_{t_1: (t_1, t_2) \in \mathcal{A}} x_{t_1 t_2} + v_{t_2} = \sum_{t_2: (t_1, t_2) \in \mathcal{A}} x_{t_1 t_2} + w_{t_1} = 1, \quad (1b)$$

$$\alpha_{t_1} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} \leq \alpha_{t_2} + M \cdot (1 - x_{t_1 t_2}) \quad (1c)$$

$$\alpha_t + \delta_{st}^{\text{school}} + \underline{\omega}_{st}^{\text{school}} \leq 5 \cdot \tau_s \leq \alpha_t + \delta_{st}^{\text{school}} + \bar{\omega}_{st}^{\text{school}}, \quad (1d)$$

$$\alpha_{t_1} + \delta_{t_1 t_2}^{\text{feeder}} + \underline{\omega}_{t_1 t_2}^{\text{change}} \leq \alpha_{t_2} + \delta_{t_1 t_2}^{\text{collector}} \leq \alpha_{t_1} + \delta_{t_1 t_2}^{\text{feeder}} + \bar{\omega}_{t_1 t_2}^{\text{change}}, \quad (1e)$$

$$\underline{\alpha}_t \leq \alpha_t \leq \bar{\alpha}_t, \underline{\tau}_s \leq 5 \cdot \tau_s \leq \bar{\tau}_s \quad (1f)$$

$$v, w \in \{0, 1\}^{|\mathcal{V}|}, x \in \{0, 1\}^{|\mathcal{A}|}, \tau \in \mathbb{Z}^{|\mathcal{S}|}, \alpha \in \mathbb{Z}^{|\mathcal{V}|}. \quad (1g)$$

3 The Primal Side

For the computation of feasible solutions we suggest an extension of the classical greedy construction methodology to what we call *parametrized greedy heuristic* (or PGreedy, for short). Greedy-type construction heuristics are used in many special-purpose optimization software packages, where feasible solutions for some combinatorial problems are required after a short amount of time. They start from scratch and step-by-step insert always the best myopic candidate while finding an answer to the given problem instance. To obtain in

fact good solutions, the crucial point within every greedy algorithm is having a proper criterion that selects these local solutions and thus is responsible for the search direction. This criterion is usually formulated in terms of a search function that assigns a certain value to each possible candidate, and the greedy algorithm then selects the candidate with the smallest value. PGreedy extends the ability of classical greedy construction heuristics by including more features into the search function, which are weighted against each other. It thus can be seen as a multicriteria extension of the greedy search function.

For the scoring function we discuss two alternatives. Similar to the classical nearest neighbor heuristic for the travelling salesman problem one might take

$$s_{t_1 t_2} := \delta_{t_1 t_2}^{\text{shift}}, \quad (2)$$

as scoring function. Hence in each step of the greedy construction heuristic an arc $(t_1, t_2) \in \mathcal{A}_k$ with minimal distance $\delta_{t_1 t_2}^{\text{shift}}$ is selected. As we will see in Section 5, this scoring function yields very poor solutions with a high number of vehicles, because it is by definition “blind” for everything that has to do with time windows. In (1) the connection of two trips t_1 and t_2 does affect some or all of the time windows. Vice versa, the time windows affect the set of deadhead trips that remain to be selected in the next round of the heuristic. Thus, we seek a scoring function that does not only take into account the time for the deadhead trip, but also takes care of the necessary changes in the corresponding time windows. If the time windows are narrowed too early in the course of the algorithms, the number of deployed buses quickly increases, because no “flexibility” remains. Thus, we introduce a scoring function that prefers those connections that do not (or at least not too much) change time windows of other trips or schools. For this, we define

$$\begin{aligned} s_{t_1 t_2}(\lambda) := & \lambda_1 \cdot \delta_{t_1 t_2}^{\text{shift}} \\ & + \lambda_2 \cdot \left| \frac{\alpha_{t_1}}{\bar{\alpha}_{t_1}} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} - \frac{\alpha_{t_2}}{\bar{\alpha}_{t_2}} \right| \\ & + \lambda_3 \cdot \left| \frac{\alpha_{t_1}}{\bar{\alpha}_{t_1}} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} - \bar{\alpha}_{t_2} \right| \\ & + \lambda_4 \cdot \left| \frac{\alpha_{t_1}}{\bar{\alpha}_{t_1}} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} - \alpha_{t_2} \right| \\ & + \lambda_5 \cdot \left| \frac{\alpha_{t_1}}{\bar{\alpha}_{t_1}} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} - \bar{\alpha}_{t_2} \right| \end{aligned} \quad (3)$$

with $\lambda = (\lambda_1, \dots, \lambda_5) \in \mathbb{Q}_+^5$. The search of a good set of parameters $\lambda \in \mathbb{Q}_+^5$ is done in an outer loop by some global optimization search scheme [1, 3].

4 The Dual Side

In order to obtain lower bounds for (1) one can use classical integer programming techniques such as branch-and-cut. Before the actual solution process starts, the solver usually enters a preprocessing phase, where it tries to obtain a more compact formulation of the problem.

One preprocessing technique that has shown to be effective in practice is the *bounds strengthening* procedure. Applied to (1), this procedure works

as follows. School and trip starting times are coupled by the minimum and maximum waiting time restrictions of inequalities (1d). The time window for the starting time of school s can be propagated onto the starting time window of trip t for all $(s, t) \in \mathcal{P}$. As an example, we obtain from (1d) and the bounds on the variables (1f):

$$\alpha_t = 5 \cdot \tau_s - \delta_{st}^{\text{school}} - \underline{\omega}_{st}^{\text{school}} \leq \bar{\tau}_s - \delta_{st}^{\text{school}} - \underline{\omega}_{st}^{\text{school}}. \quad (4)$$

We now compare the right-hand side of (4) with the previous upper bound $\bar{\alpha}_t$ on α_t . If it is less, a new *improved* upper bound is found. In general, we set

$$\bar{\alpha}_t := \min\{\bar{\alpha}_t, \bar{\tau}_s - \delta_{st}^{\text{school}} - \underline{\omega}_{st}^{\text{school}}\}, \quad (5)$$

$$\underline{\alpha}_t := \max\{\underline{\alpha}_t, \underline{\tau}_s - \delta_{st}^{\text{school}} - \bar{\omega}_{st}^{\text{school}}\}. \quad (6)$$

Vice versa, the trip time window can be propagated onto the school time window. This procedure is today standard in most branch-and-cut solvers. However, only one inequality is used after the other to improve the bounds.

Much stronger results can be obtained by *higher order bounds strengthening*, if more than one inequality is taken into consideration (two or three, for example): Consider two trips $(t_1, t_2) \in \mathcal{A}$ with $(t_1, t_2) \in \mathcal{C}$, and the following subsystem of (1):

$$\begin{aligned} \alpha_{t_1} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} + \underline{\omega}_{t_1 t_2}^{\text{idle}} - M \cdot (1 - x_{t_1 t_2}) &\leq \alpha_{t_2}, \\ \alpha_{t_1} + \delta_{t_1 t_2}^{\text{feeder}} + \bar{\omega}_{t_1 t_2}^{\text{change}} &\geq \alpha_{t_2} + \delta_{t_1 t_2}^{\text{collector}}. \end{aligned} \quad (7)$$

By adding up these two inequalities we can eliminate the variables $\alpha_{t_1}, \alpha_{t_2}$:

$$\delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} + \underline{\omega}_{t_1 t_2}^{\text{idle}} - \delta_{t_1 t_2}^{\text{feeder}} - \bar{\omega}_{t_1 t_2}^{\text{change}} + \delta_{t_1 t_2}^{\text{collector}} \leq M \cdot (1 - x_{t_1 t_2}). \quad (8)$$

Hence if $\delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} + \underline{\omega}_{t_1 t_2}^{\text{idle}} - \delta_{t_1 t_2}^{\text{feeder}} - \bar{\omega}_{t_1 t_2}^{\text{change}} + \delta_{t_1 t_2}^{\text{collector}} > 0$ we set $x_{t_1 t_2} := 0$ and thus eliminate this decision variable.

Consider two trips $(t_1, t_2) \in \mathcal{A}$ with $(s, t_1), (s, t_2) \in \mathcal{P}$ for some school $s \in \mathcal{S}$, and the following subsystem of (1):

$$\begin{aligned} \alpha_{t_1} + \delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} + \underline{\omega}_{t_1 t_2}^{\text{idle}} - M \cdot (1 - x_{t_1 t_2}) &\leq \alpha_{t_2}, \\ \alpha_{t_1} + \delta_{st_1}^{\text{school}} + \bar{\omega}_{st_1}^{\text{school}} &\geq 5 \cdot \tau_s, \\ \alpha_{t_2} + \delta_{st_2}^{\text{school}} + \underline{\omega}_{st_2}^{\text{school}} &\leq 5 \cdot \tau_s. \end{aligned} \quad (9)$$

By adding up these three inequalities we can eliminate the variables $\alpha_{t_1}, \alpha_{t_2}, \tau_s$:

$$\delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} + \underline{\omega}_{t_1 t_2}^{\text{idle}} - \delta_{st_1}^{\text{school}} - \bar{\omega}_{st_1}^{\text{school}} + \delta_{st_2}^{\text{school}} + \underline{\omega}_{st_2}^{\text{school}} \leq M \cdot (1 - x_{t_1 t_2}). \quad (10)$$

Thus if $\delta_{t_1}^{\text{trip}} + \delta_{t_1 t_2}^{\text{shift}} + \underline{\omega}_{t_1 t_2}^{\text{idle}} - \delta_{st_1}^{\text{school}} - \bar{\omega}_{st_1}^{\text{school}} + \delta_{st_2}^{\text{school}} + \underline{\omega}_{st_2}^{\text{school}} > 0$ we set $x_{t_1 t_2} := 0$.

5 Computational Results

The greedy and PGreedy heuristics were implemented in C++ on a 2.6GHz Intel-Pentium IV computer running debian Linux [4] and tested with five real-world data sets from different counties in Mecklenburg-Western Pomerania,

North Rhine-Westphalia, and Saxony-Anhalt (see first or left block of Table 1). For the dual bounds of the integer programs we used the ILOG CPLEX9 solver.

The sizes of sets in the instances are shown in the second block of Table 1. The objective function for the current school starting times and bus schedules is given in the third block. A comparison of the solutions found by the greedy and the PGreedy heuristic is given in the forth block. PGreedy solutions are able to reduce the number of deployed buses by 10 – 30% (note that a single bus is worth about 35,000 Euro per year). In the last block of Table 1 we show the dual bounds, first with CPLEX standard presolve (cpx-std), and then with higher order bounds strengthening (cpx-hobs). In three out of five cases the dual bound has improved due to this new presolving technique. In one case (Wernigerode), the solver was able to find a global optimal solution after 1,600 seconds. (For all other primal and dual computations the time limit was set to 3,600 seconds.)

Table 1. Size of the input data sets, primal solutions, and dual bounds

instance	$ \mathcal{V} $	$ \mathcal{A} $	$ \mathcal{S} $	$ \mathcal{P} $	$ \mathcal{C} $	current	greedy	PGreedy	cpx-std	cpx-hobs
Demmin	247	60,762	43	195	165	82.0423	81.0358	65.0882	39.1513	39.1369
Steinfurt	490	239,610	102	574	406	226.0711	227.0556	175.1595	88.9567	88.4651
Soest	191	36,290	82	294	182	90.0488	83.0410	66.1106	20.1343	21.5686
W'rode	134	17,822	37	201	204	43.0093	49.0138	38.0268	37.3984	38.0198
Guetersloh	404	162,812	84	579	708	176.0725	180.0663	134.1704	85.6738	86.1151

References

1. A. Fügenschuh: Parametrized Greedy Heuristics in Theory and Practice. Lecture Notes in Computer Science 3636: 21 – 31, 2005.
2. A. Fügenschuh: The Integrated Optimization of School Starting Times and Public Bus Services. PhD Thesis, Logos Verlag Berlin, 165 pages, 2005.
3. A. Fügenschuh: The Vehicle Routing Problem with Coupled Time Windows. Central European Journal of Operations Research 14(2), 2006.
4. A. Fügenschuh, A. Martin: Verfahren und Vorrichtung zur automatischen Optimierung von Schulanfangszeiten und des öffentlichen Personenverkehrs und entsprechendes Computerprogramm. Deutsches Patent 10 2004 020 786.0, 2004.
5. A. Fügenschuh, A. Martin: A Multicriteria Approach for Optimizing Bus Schedules and School Starting Times. To appear in Annals of Operations Research.
6. A. Fügenschuh, P. Stöveken: Integrierte Optimierung der Schulanfangszeit und des Nahverkehrs-Angebots. Heureka'05 Proceedings, Forschungsgesellschaft für Straßen- und Verkehrswesen, Köln, 265 – 278, 2005. Also published in: Straßenverkehrstechnik 49(6): 281 – 287, 2005 (In German).
7. P. Stöveken: Wirtschaftlicherer Schulverkehr: ÖPNV-Optimierung mit erfolgsabhängiger Honorierung. Der Nahverkehr 3: 65 – 68, 2000 (In German).

Dynamisches Bestandsmanagement in der Kreislauflogistik

Rainer Kleber

Fakultät für Wirtschaftswissenschaft, Otto-von-Guericke Universität Magdeburg,
Germany

rainer.kleber@ww.uni-magdeburg.de

Summary. Quantitative Ansätze zum Bestandsmanagement im Rahmen der Kreislauflogistik fokussieren hauptsächlich auf Losgrößen- und Sicherheitsbestände. Aufgrund der dabei genutzten statischen Modellannahmen sind sie kaum in der Lage, die häufig in der Praxis vorzufindenden hohen Bestände insbesondere an Altprodukten zu erklären. Eine explizite Berücksichtigung dynamischer Einflüsse, wie sie beispielsweise Saisonalitäten, Produktlebenszyklen oder auch die Kostendynamik darstellen, führt zu neuen Motiven für die Lagerhaltung. Aufgabe der Dissertation [5] war es, solche Motivationen zu identifizieren. Dabei wurde auf eine zeitstetige Modellierung zurückgegriffen. Als Lösungsmethodik wurde Pontryagins Maximumprinzip genutzt, mit welchem generelle Struktureigenschaften optimaler Lösungen für ganze Problemklassen ermittelt werden können. Dieser Artikel gibt einen Überblick über wesentliche Resultate der Dissertation.

1 Einleitung

Während sich Hersteller von Gebrauchsgütern in der Vergangenheit auf eine möglichst effiziente Produktion und Distribution ihrer Erzeugnisse konzentrieren konnten, sehen sie sich heute in zunehmendem Maße einem Verantwortungsbereich gegenüber, der über die Nutzungsphase der Güter hinausgeht. Die Entsorgung von Altprodukten, früher eine öffentliche Domäne, muss nun zunehmend im Planungsprozess berücksichtigt werden. Dies führt konsequenterweise zu zusätzlichen Einschränkungen, allerdings auch zu neu zu erschließenden Kostensenkungspotentialen, sofern die Hersteller neben der Verantwortung auch den Zugriff auf Altprodukte erhalten. Findet nun generell eine Altproduktrücknahme durch den Hersteller statt, lässt sich durch Recycling von Materialien der Bedarf an immer teureren Rohstoffen senken oder es können Nachfragen beispielsweise nach Ersatzteilen durch die Aufarbeitung von Altprodukten gedeckt werden. Zu beantwortende Fragen in diesem Zusammenhang beinhalten:

1. Welche Entsorgungsoption sollte gewählt werden (z.B. Beseitigung, Recycling oder Aufarbeitung)?
2. Wie sollten Neuproduktion und Aufarbeitung koordiniert werden, und wie sollte mit Altprodukten umgegangen werden, die nicht unmittelbar genutzt werden können?

Die erste eher strategische Entscheidung beeinflusst das Maß der Werterhaltung und die Entsorgungskosten. Von besonderem Interesse ist hier die Aufarbeitung, da diese genutzt werden kann, um Teile der Neuproduktion zu ersetzen. Die Spannbreite aufgearbeiteter Produkte reicht von Chemikalien in der pharmazeutischen Industrie über Einwegkameras bis hin zu komplexen Produkten wie Motoren in der Automobilindustrie oder Kopierautomaten. Die zweite operative Fragestellung beinhaltet unter Anderem eine Abwägung zwischen Lagerhaltungskosten und den Kosteneinsparungen der Nutzung unterschiedlicher Entsorgungsoptionen.

Grundsätzlich führt die Berücksichtigung von Altprodukten zu einer höheren Komplexität der logistischen Prozesse, da nun Distribution (vorwärts) und Redistribution (rückwärts gerichtet) miteinander koordiniert werden müssen. Die Berücksichtigung der Kunden als Akteure im Rahmen der Redistribution bringt zusätzliche Unsicherheiten mit sich. Der Planungsprozess stellt sich komplizierter dar, da nun zwei Quellen zur Bedarfsdeckung und weitere Optionen für die Bestandshaltung existieren. Diesen Herausforderungen stellt sich das Forschungsgebiet der Kreislauflogistik (Reverse Logistics). Neben einem erweiterten (Closed Loop) Supply Chain Mangement und der Redistribution stellt hier das integrierte Produktions- und Bestandsmanagement einen wesentlichen Forschungsbereich dar [1]. Dieser Artikel wie auch die Dissertation [5] befassen sich mit dem letztgenannten Bereich.

Im Rahmen des Bestandsmanagements lassen sich in Abhängigkeit von ihrer jeweiligen Motivation verschiedene Bestandsarten unterscheiden. Sicherheitsbestände werden im hier betrachteten Zusammenhang genutzt, um kurzfristige Unsicherheiten in der Nachfrage und in der Verfügbarkeit von Altprodukten auszugleichen; Losgrößenbestände dienen dem Ausgleich von Lagerhaltungs- und auftragsabhängigen Kosten. Obwohl beide Motivationen bereits hinreichend Beachtung in der Literatur gefunden haben [11, 10], sind sie kaum in der Lage, tatsächlich in der Praxis vorzufindende hohe Bestände zu erklären. Zumeist unterstellen die genutzten Modelle stationäre Bedingungen und vernachlässigen dynamische Einflüsse wie zeitabhängige Nachfragen, Rückflüsse, Kapazitäten oder auch Kosten. Bei expliziter Berücksichtigung dieser Effekte lassen sich neue Motivationen für Lagerhaltung in Systemen der Kreislauflogistik isolieren, von denen einige im Folgenden vorgestellt werden sollen. Hierbei wird ein deterministisches einstufiges Einproduktsystem (wie in Abbildung 1 dargestellt) mit einer linearen Kostenstruktur betrachtet, in welchem die Aufarbeitung von Altprodukten einen direkten Vorteil gegenüber der Neuproduktion besitzt.

Da aufgrund des eher mittel- bis langfristigen Planungshorizonts die Modellierung in stetiger Zeit erfolgte, wurde als Lösungsmethodik Pontryagins Maximumprinzip angewendet. Diese Vorgehensweise bietet, im Gegensatz beispielsweise zur linearen Optimierung, den Vorteil, dass generelle Struktureigenschaften optimaler Lösungen für ganze Problemklassen ermittelt werden können.

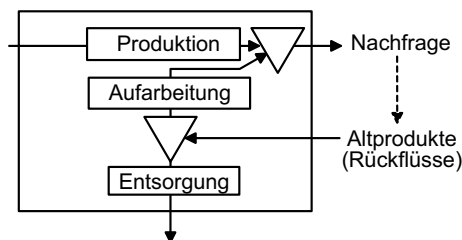


Fig. 1. Logistisches System mit Altproduktrücknahme.

2 Operative Aspekte

In einem Basismodell (vergleiche auch [9]) konnte nachgewiesen werden, dass die Lagerung von Altprodukten genau dann sinnvoll ist, wenn auf Perioden mit Altproduktüberschuss solche mit Nachfrageüberschuss folgen (siehe Abbildung 2). Die verfügbaren Rückflüsse können hier als Aufarbeitungskapazität interpretiert werden, die durch einen antizipativen Bestandsaufbau dynamisch erhöht werden kann. Da durch die Lagerhaltung ebenfalls Kosten verursacht werden und der Wert der Altprodukte mit fortschreitender Zeit sinkt, ist dies nur für eine maximale Lagerdauer sinnvoll.

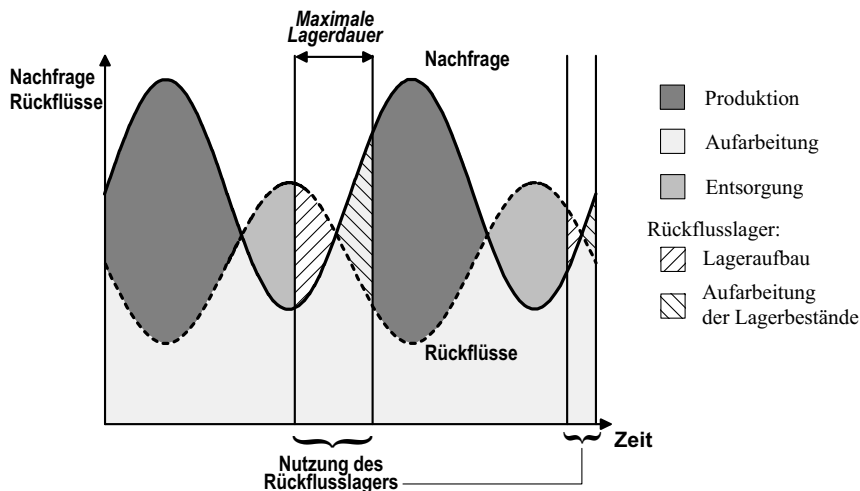


Fig. 2. Nutzung von Antizipationsbeständen im Basismodell.

Bei Einführung von exogenen Kapazitätsbeschränkungen in Form maximaler Produktions- bzw. Aufarbeitungsmengen lassen sich vielfältige Situationen mit Lagerhaltung identifizieren, von denen hier jedoch nur einige angerissen werden können. Bei beschränkter Produktionskapazität sind nunmehr beide Quellen limitiert; ohne Aufarbeitung müsste stets dann ein Saisonbestand im Fertigproduktlager aufgebaut werden, sobald die Restriktion bindend wird. Unter Berücksichtigung der

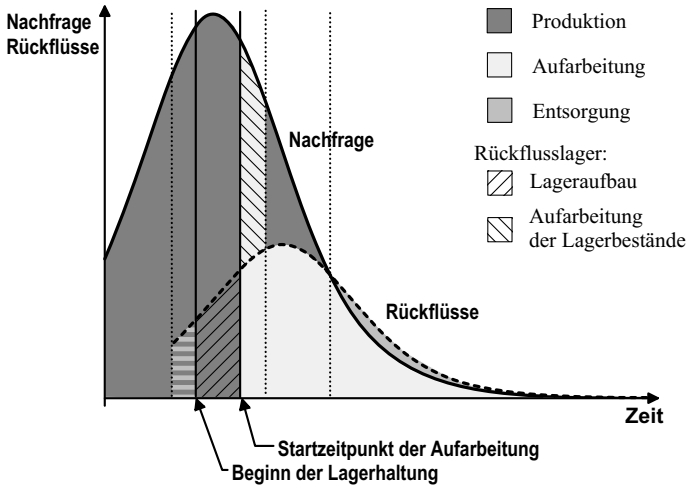


Fig. 3. Startzeitpunkt der Lagerhaltung und der Aufarbeitung bei Investitionsentscheidungen unter dynamischen Rahmenbedingungen.

Nutzung von Altprodukten wird eine Bestandshaltung nötig, sobald die Nachfrage die augenblickliche Gesamtkapazität des Systems (bestehend aus maximaler Produktionsmenge und Anfall an Altprodukten) übersteigt. Wenn Lagerhaltung im Rückflusslager günstiger und die Aufarbeitung unbeschränkt ist, wird der benötigte Engpassbestand vorzugsweise im Lager für Altprodukte aufgebaut. Die Lagerdauer ist hier zum einen von der Engpasshöhe aber auch vom Anfall an Altprodukten abhängig. Aufgrund dessen kann kurzzeitig auch das Fertigproduktlager genutzt werden, um einen schnelleren Lageraufbau im Altproduktlager zu ermöglichen. Bei eingeschränkter Aufarbeitungskapazität könnte die Nachfrage eigentlich zu jedem Zeitpunkt auch aus laufender (unbeschränkter) Produktion befriedigt werden. Aber auch hier kann in Analogie zum Antizipationsbestand ein Bestandsaufbau im Lager für Altprodukte erfolgen, um knappe Aufarbeitungskapazitäten besser auszulasten. Darüber hinaus kann auch das Fertigproduktlager genutzt werden, um diesen Effekt zu verstärken.

3 Strategische Aspekte

Eine wesentliche strategische Fragestellung beinhaltet die Entscheidung, ob sich Aufarbeitung grundsätzlich lohnt. Dagegen sprechen könnten die Notwendigkeit eines aufwendigeren "Mehrweg" - Produktdesigns, zusätzliche Investitionen in Produktionsanlagen und Investitionen in Aufarbeitungsanlagen. Unter dynamischen Rahmenbedingungen, bestehend aus einem Produktlebenszyklus und dessen Äquivalent für Altprodukte (siehe Abbildung 3), führt dies zu der Frage, wann Aufarbeitungsaktivitäten gestartet werden sollten und welchen Einfluss eine Lagerhaltung von Altprodukten auf diese Entscheidung hat. Neben der Analyse des zugrundeliegenden dynamischen Entscheidungsproblems konnte auf Basis einer umfangreichen numerischen Untersuchung [6] festgestellt werden, dass die gezielte Wahl

des Startzeitpunktes der Aufarbeitung und eine effiziente antizipative Lagerhaltung einen substantiellen Einfluss auf die Vorteilhaftigkeit der Aufarbeitungsoption haben.

Aufarbeitungsprozesse sind üblicherweise arbeitsintensiv und somit Lerneffekten unterworfen. Eine Ausnutzung dieser Effekte kann dazu führen, dass eine anfänglich nicht vorteilhafte Aufarbeitung aufgrund eines hinreichend hohen Volumens insgesamt zur lohnenswerten Option wird. Aus strategischer Sicht ergeben sich ähnliche Resultate wie bei der Berücksichtigung von Investitionen. Insbesondere kann die Aufarbeitung auch hier später als zu Beginn des Planungshorizontes starten und bei hohem anfänglichem Aufarbeitungsnachteil kann ein Bestandsaufbau im Rückfluslager zum weiteren Aufschieben des Startzeitpunktes führen.

4 Erweiterungen

Es existieren verschiedenste Möglichkeiten zur Erweiterung und Anwendung des Modellansatzes (vergleiche auch [4]). Ein absichtliches Vormerken von Teilen der Nachfrage (‐Negative Antizipationsbestände‐, siehe [3]) erlaubt es unter geeigneten Bedingungen, weitere Nachfragen durch Aufarbeitung anstelle von Produktion zu befriedigen. Die bedarfsgerechte Akquisition der Altprodukte durch eine entsprechende dynamische Gestaltung des Rücknahmepreises führt zum Abwälzen von Lagerhaltungskosten auf den Konsumenten [8]. Andere Erweiterungen stellen die Berücksichtigung mehrerer Aufarbeitungsoptionen [7] sowie von Vorlaufzeiten bei der Produktion und Aufarbeitung [2] dar.

References

1. Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (2004) Quantitative models for reverse logistics decision management. In: Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (Hrsg.) *Reverse Logistics*, S. 25-41. Springer, Berlin.
2. Kiesmüller GP (2003) Optimal control of a one product recovery system with leadtimes. *International Journal of Production Economics*, 81-82:333-340.
3. Kiesmüller GP, Minner S, Kleber R (2000). Optimal control of a one product recovery system with backlogging. *IMA Journal of Mathematics Applied in Business and Industry*, 11:189-207.
4. Kiesmüller GP, Minner S, Kleber R (2004). Managing dynamic product recovery: An optimal control perspective. In: Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (Hrsg.) *Reverse Logistics*, S. 221-247. Springer, Berlin.
5. Kleber R (2006) *Dynamic Inventory Management in Reverse Logistics*. Springer, Berlin.
6. Kleber R (2006) The integral decision on production/remanufacturing technology and investment time in product recovery. *OR Spectrum* 28: 21-51.
7. Kleber R, Minner S, Kiesmüller GP (2002). A continuous time inventory model for a product recovery system with multiple options. *International Journal of Production Economics*, 79:121-141.

8. Minner S, Kiesmüller GP (2002). Dynamic product acquisition in closed loop supply chains. FEMM Working Paper 9/2002, Otto-von-Guericke-Universität Magdeburg.
9. Minner S, Kleber R (2001) Optimal control of production and remanufacturing in a simple recovery model with linear cost functions. *OR Spektrum* 23:3-24.
10. Minner S, Lindner G (2004) Lot sizing decisions in product recovery management. In: Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (Hrsg.) *Reverse Logistics*, S. 157-179. Springer, Berlin.
11. van der Laan EA, Kiesmüller GP, Kuik R, Vlachos D, Dekker R (2004) Stochastic inventory control for product recovery management. In: Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN (Hrsg.) *Reverse Logistics*, S. 181-220. Springer, Berlin.

Periodic Timetable Optimization in Public Transport^{*}

Christian Liebchen

Institut für Mathematik, Kombinatorische Optimierung und Graphenalgorithmien,
Technische Universität Berlin, Germany
`liebchen@math.tu-berlin.de`

Summary. „The timetable is the essence of the service offered by any provider of public transport.“ (Jonothan Tyler, CASPT 2006)

Despite this observation, in the practice of planning public transportation, only some months ago OR decision support has still been limited to operations planning (vehicle scheduling, duty scheduling, crew rostering). We describe the optimization techniques that were employed in computing the very first optimized timetable that went into daily service: the 2005 timetable of Berlin Underground. This timetables improved on both, the passenger travel times and the operating efficiency of the company.

The basic graph model, the Periodic Event Scheduling Problem (PESP), is known for 15 years and it had attracted many research groups. Nevertheless, we report on significant progress that has been made only recently on issues like solution strategies or modeling capabilities. The latter even includes the integration of further planning tasks in public transport, such as line planning.

On the theory side, we give a more precise notion of the asymptotical complexity of the PESP, by providing a MAXSNP-hardness proof as a kind of negative result. On the positive side, the design of more efficient algorithms gave rise to a much deeper understanding of cycle bases of graphs, another very hot topic in discrete mathematics during the last three years. In 2005, this culminated in both, drawing the complete map for the seven relevant classes of cycle bases, and the design of the fastest algorithms for the Minimum Directed Cycle Basis Problem and for the Minimum 2-Basis Problem.

The book version of this extended abstract is available as reference [8].

1 Timetabling

It is a very important competitive advantage of public transport to be much less expensive than a taxi service. This requires many passengers to share the same vehicle. Typically, this is achieved by offering public transport along fixed sets of routes, the lines. These serve as input to timetabling.

^{*} This work has been supported by the DFG Research Center MATHEON in Berlin.

There is a large toolbox of different types of timetables, which we introduce from the most general one to the most specialized one:

- timetables that are composed of individual trips,
- periodic timetables, i.e. the headway between any two successive trips of the same line is the same,
- symmetric periodic timetables, and
- so-called “Integrated Fixed-Interval Timetables.”

Here, a periodic timetable is called *symmetric*, if for every passenger the transfer times that he faces during his outbound trip are identical to the transfer times during his return trip, which here is assumed to have the same route. In particular, the periodic timetables of most European national railway companies are indeed symmetric, because marketing departments consider this being a competitive advantage—at least in long-distance traffic.

Theorem 1 ([7]). *There exist example networks showing that each more specialized family of timetables causes a nominal loss in quantifiable criteria, such as average passenger waiting time.*

We are only aware of periodic timetables being able to clearly outweigh their nominal loss (when comparing with general irregular timetables) by adding benefit in qualitative criteria. Hence, in the remainder we focus on periodic timetables.

Typically, the period time T varies over the day. For instance, Berlin Underground distinguishes

- rush hour service ($T = 4$ minutes),
- „normal“ service ($T = 5$ minutes),
- weak traffic service ($T = 10$ minutes, when retail shops are closed), and
- night service ($T = 15$ minutes, only on weekends).

Computing “the timetable” thus decomposes into computing a periodic timetable for each period time, and finally glue these together.

2 A Model for Periodic Timetabling

A literature review of different models for periodic scheduling reveals that the most promising earlier studies on medium-sized networks are based on the Periodic Event Scheduling Problem (PESP, [18]), see [17, 14, 9, 16]. The vertices in this graph model represent events, where an event $v \in V$ is either an arrival or a departure of a directed line in a specific station. A timetable π assigns to each vertex v a point in time $\pi_v \in [0, T)$ within the period time T . Constraints may then be given in the following form.

T-PERIODIC EVENT SCHEDULING PROBLEM (T-PESP)

Instance: A directed graph $D = (V, A)$ and vectors $\ell, u \in \mathbb{Q}^A$.

Task: Find a vector $\pi \in [0, T)^V$ that fulfills

$$(\pi_v - \pi_u - \ell_a) \bmod T \leq u_a - \ell_a \quad (1)$$

(or $\pi_v - \pi_u \in [\ell_a, u_a]_T$, for short) for every arc $a = (u, v) \in A$, or decide that none exists.

In Figure 1 we provide an example instance of T -PESP, which contains the events of two pairs of directed lines and two stations.

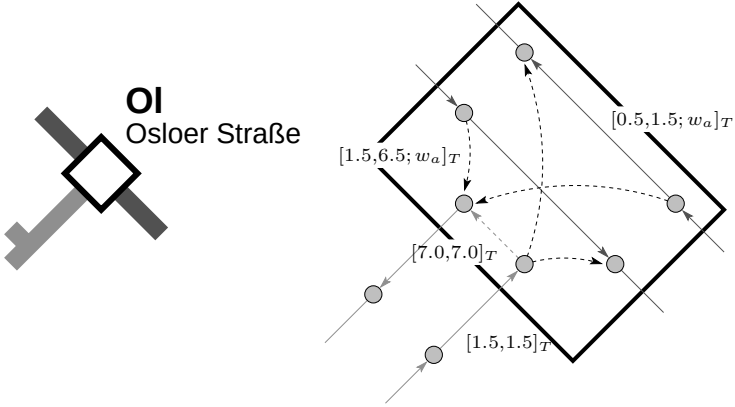


Fig. 1. A PESP model for two lines and two stations

Here, the straight arcs model either stops (within the black box that represents the station) or trips, and the dotted arcs model either passenger transfers or turnarounds of the trains. Besides these most elementary requirements, there have been modeled most practical requirements that railway engineers have ([11]). This even includes decisions of line planning and vehicle scheduling, which traditionally were treated as fully separate planning steps ([1]).

Unfortunately, this modeling power has its price in terms of complexity.

Theorem 2. *Let a set of PESP constraints be given. Finding a timetable vector π that satisfies a maximum number of constraints is MAXSNP-hard.*

To make T -PESP accessible for integer programming (IP) techniques, the modulo-operator is resolved by introducing integer variables:

$$\begin{aligned}
 \min \quad & w^T(B^T\pi + Tp) \\
 \text{s.t.} \quad & B^T\pi + Tp \leq u \\
 & B^T\pi + Tp \geq \ell \\
 & \pi \in [0, T)^V \\
 & p \in \{0, 1, 2\}^A.
 \end{aligned} \tag{2}$$

Here, the matrix B is the vertex-arc incidence matrix of the constraint graph $D = (V, A)$.

But this is not the only way to formulate T -PESP as an IP. Rather, we may replace the vertex variables π (or node potentials) — which carry time information — with arc variables x (tensions), and/or replace the integer arc variables p with integer cycle variables z . This way, we end with the following integer program ([14])

$$\begin{aligned}
 \min \quad & w^T x \\
 \text{s.t.} \quad & x \leq u \\
 & x \geq \ell \\
 & \Gamma^T x - Tz = 0 \\
 & x \in \mathbb{Q}^A \\
 & z \in \mathbb{Z}^B,
 \end{aligned} \tag{3}$$

where Γ denotes the arc-cycle incidence matrix of an integral cycle basis B of the constraint graph $D = (V, A)$. There have also been identified lower and upper bounds on these integer variables z .

Theorem 3 ([15]). *Let C be an oriented circuit and z_C the integer variable that we associate with it. The following inequalities are valid*

$$\left\lfloor \frac{1}{T} \left(\sum_{a \in C^+} \ell_a - \sum_{a \in C^-} u_a \right) \right\rfloor =: \underline{z}_C \leq z_C \leq \bar{z}_C := \left\lceil \frac{1}{T} \left(\sum_{a \in C^+} u_a - \sum_{a \in C^-} \ell_a \right) \right\rceil.$$

The following rule-of-thumb could be derived from empirical studies.

Remark 1 ([4]). The shorter a circuit $C \in B$ with respect to the sum of the spans $u_a - \ell_a$ of its arcs, the less integer values the corresponding variable z_C may take. Moreover, the less values all the integer variables may take, the shorter the solution times for solving this IP. $\square$

3 Integral Cycle Bases

In a directed graph $D = (V, A)$, we consider *oriented circuits*. These consist of forward arcs and maybe also backward arcs, such that re-orienting the backward arcs yields a directed circuit. The *incidence vector* $\gamma_C \in \{-1, 0, 1\}^A$ of an oriented circuit C has a plus (minus) one entry precisely for the forward (backward) arcs of C . Then, the *cycle space* $\mathcal{C}(D)$ can be defined as

$$\mathcal{C}(D) := \text{span}(\{\gamma_C \mid C \text{ is an oriented circuit of } D\}).$$

A *cycle basis* B of $\mathcal{C}(D)$ is a set of oriented circuits, which is a basis of $\mathcal{C}(D)$. An *integral cycle basis* allows to combine every oriented circuit of D as an *integer* linear combination of the basic circuits.

Fortunately, in order to decide upon the integrality of a cycle basis, we do not have to check all these linear combinations.

Lemma 1 ([4]). *Let Γ be the arc-cycle incidence matrix of a cycle basis. For two submatrices Γ_1, Γ_2 with $\text{rank}(\Gamma_1) = \text{rank}(\Gamma_2) = \text{rank}(\Gamma)$, there holds*

$$\det \Gamma_1 = \pm \det \Gamma_2. \quad (4)$$

Definition 1 (Determinant of a cycle basis, [4]). *Let B be a cycle basis and Γ_1 as in the above lemma. We define the determinant of B as*

$$\det B := |\det \Gamma_1|. \quad (5)$$

Theorem 4 ([4]). *A cycle basis B is integral, if and only if $\det B = 1$.*

According to Remark 1, in the application of periodic timetabling we seek for a minimum integral cycle basis of D . To illustrate the benefit of short integral cycle we provide the following example.

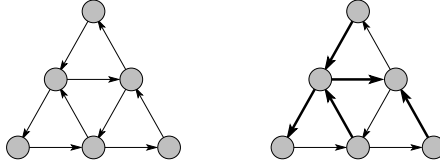


Fig. 2. The sunflower graph $SF(3)$, and a spanning tree

Example 1. Consider the sunflower graph $SF(3)$ in Figure 2. Assume each arc models a PESP constraint of the form $[7,13]_{10}$, i.e. subject to a period time of $T = 10$.

According to the initial IP formulation (2), we could deduce that in order to identify an optimum timetable simply by considering every possible timetable vector π we need to check $|\{0, \dots, 9\}|^{|V|} = 1,000,000$ vectors π . Alternatively, we might check for $|\{0,1,2\}|^{|A|} = 19,683$ vectors p to pervade the search space. It will turn out, that these two perspectives have much redundancies within them.

In contrast, the valid inequalities of Theorem 3 reveal that every 4-circuit C in D yields an integer variable z_C which may only take the three values $\{-1,0,1\}$. Even better, a triangle C in D induces a fixed variable z_C .

Thus, the integral cycle basis B that can be derived from the spanning tree F (Fig. 2 on the right) already reduces the upper bound on the size of the search space to only $1 \cdot 1 \cdot 1 \cdot 3 = 3$ possible vectors for z . Moreover, considering the minimum cycle basis of D — which for this graph turns out to be integral as it consists of the circuits that bound the four finite faces of this plane graph — we end with just one single vector z describing the complete instance. $\square$

Ideally, we would like to compute a minimum integral cycle basis of D according to the edge weights $u_a - \ell_a$. Unfortunately, we are not aware of the asymptotical complexity of this combinatorial optimization problem. However, recently there has been achieved much progress on the the asymptotical complexity for the corresponding minimum cycle basis problems for related classes of cycle bases, see [13, 2] and referenced therein. We depict these results in Figure 3. Notice that any of these classes demands for specific algorithms, because none of these problems coincide ([13]).

4 Summary of Computational Results

Earlier, in several autonomous computational studies, there had been applied various algorithms to periodic timetabling. We have executed the first unified computational study, which covers algorithms as variegated as a Cut-and-Branch Algorithm for Integer Programs, Constraint Programming, and even Genetic Algorithms ([12]). To any of these, quite a number of different promising parameter settings was applied. In particular for Integer Programming this amounts to hundreds of different runs on five different data sets. All the data sets have been provided to us by industrial partners, and they range from long-distance traffic over regional traffic down to undergrounds, comprising between 10 and 40 commercial lines each.

With respect to both solution quality and independence of parameter settings, both our Genetic Algorithm (GA) and our Cut-and-Branch Algorithm—which is

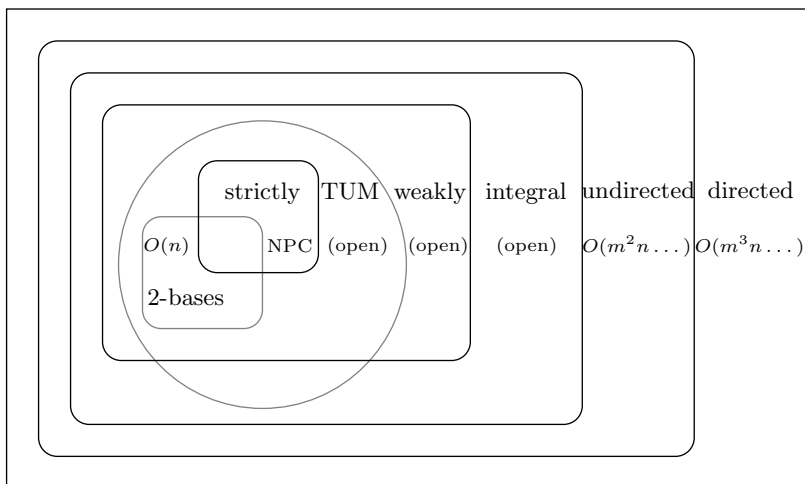


Fig. 3. Map of the complexity of the seven variants of the MINIMUM CYCLE BASIS PROBLEM for general graphs ([4, 13])

using CPLEX[®] 9.1—perform considerably well. On the one hand, IP techniques turn out to be extremely sensitive with respect to the choice of several important parameters. On the other hand, in particular for medium-sized instances for which IP techniques still attain an optimum solution, the quality achieved by the GA is somewhat worse.

5 Improvements for Berlin Underground

Most important, in a long-term cooperation with Berlin Underground we continuously kept on improving our mathematical models of the real world ([10, 6]). Finally, in 2004 we were able to formulate a mathematical program which covered *all* the practical requirements that the practitioners have. As a consequence, the optimum solution that was computed by our algorithms convinced Berlin Underground: By December 12, 2004, our timetable became the first optimized timetable that went into service—presumably worldwide. This may be compared to the fact that only in operations planning (vehicle scheduling, duty scheduling), Operations Research had already entered the practice.

Compared to the former timetable, with our timetable the passengers of Berlin Underground are offered simultaneously improvements in two key criteria, which typically are conflicting: transfer waiting time and dwell time of trains. In more detail, our achievements are:

- The number of transfers, for which a maximum transfer waiting time of 5 minutes can be guaranteed, increases from 95 to 103 (+8%).
- The maximum dwell time of any train in the network was reduced from 3.5 minutes to only 2.5 minutes (−30%).
- The timetable could even be operated with one train less.

The part of the network in which the most significant improvements have been achieved is given in Figure 4. *Network waiting time charts* emerged as a by-product from our cooperation with Berlin Underground ([6, 7]). Such charts constitute the first visualization of the transfer quality of a timetable. In particular, they made the discussion of pros and cons of different timetables most efficient, as for instance long transfer waiting times (marked in black) along important transfers (marked as **bold** arcs) become obvious.

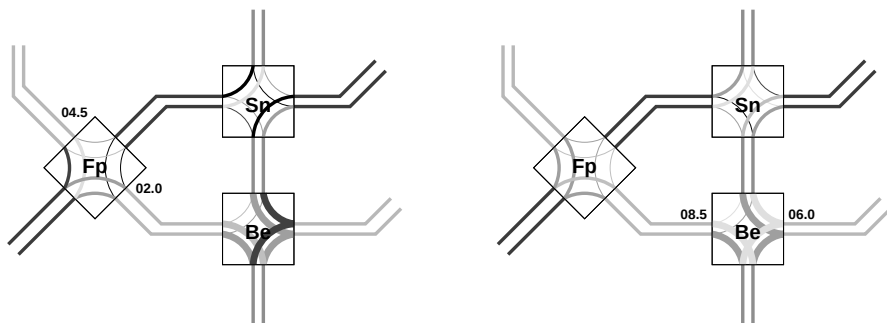


Fig. 4. Network waiting time charts for an excerpt of the Berlin subway network—before and after invoking mathematical optimization

The successful transfer from theory to practice has even been reflected by articles and interviews in nationwide newspapers and radio transmissions:

- Berliner Zeitung, November 9, 2005, in German ([3])
<http://www.berlinonline.de/berliner-zeitung/archiv/.bin/dump.fcgi/2005/1109/wissenschaft/0002/index.html>
- Deutschlandfunk, December 9, 2005, 16:35h, in German
<http://www.dradio.de/dlf/sendungen/forschak/446751/>

References

1. Michael R. Bussieck, Thomas Winter, and Uwe Zimmermann. Discrete optimization in public rail transport. *Mathematical Programming B*, 79:415–444, 1997.
2. Ramesh Hariharan, Telikepalli Kavitha, and Kurt Mehlhorn. A Faster Deterministic Algorithm for Minimum Cycle Bases in Directed Graphs. In Michele Bugliesi et al., editors, *ICALP*, volume 4051 of *Lecture Notes in Computer Science*, pages 250–261. Springer, 2006.
3. Reinhard Huschke. Schneller Umsteigen. *Berliner Zeitung*, 61(262):12, 2005. Wednesday, November 9, 2005, In German.
4. Christian Liebchen. Finding short integral cycle bases for cyclic timetabling. In Giuseppe Di Battista and Uri Zwick, editors, *ESA*, volume 2832 of *Lecture Notes in Computer Science*, pages 715–726. Springer, 2003.
5. Christian Liebchen. A cut-based heuristic to produce almost feasible periodic railway timetables. In Sotiris E. Nikolettseas, editor, *WEA*, volume 3503 of *Lecture Notes in Computer Science*, pages 354–366. Springer, 2005.

6. Christian Liebchen. Der Berliner U-Bahn Fahrplan 2005 – Realisierung eines mathematisch optimierten Angebotskonzeptes. In *HEUREKA '05: Optimierung in Transport und Verkehr, Tagungsbericht*, number 002/81. FGSV Verlag, 2005. In German.
7. Christian Liebchen. Fahrplanoptimierung im Personenverkehr—Muss es immer ITF sein? *Eisenbahntechnische Rundschau*, 54(11):689–702, 2005. In German.
8. Christian Liebchen. *Periodic Timetable Optimization in Public Transport*. dissertation.de, 2006. PhD thesis.
9. Thomas Lindner. *Train Schedule Optimization in Public Rail Transport*. Ph.D. thesis, Technische Universität Braunschweig, 2000.
10. Christian Liebchen and Rolf H. Möhring. A case study in periodic timetabling. *Electr. Notes in Theoretical Computer Science*, 66(6), 2002.
11. Christian Liebchen and Rolf H. Möhring. The modeling power of the periodic event scheduling problem: Railway timetables – and beyond. Preprint 020/2004, TU Berlin, Mathematical Institute, 2004. To appear in Springer LNCS Volume Algorithmic Methods for Railway Optimization.
12. Christian Liebchen, Mark Proksch, and Frank H. Wagner. Performance of algorithms for periodic timetable optimization. To appear in Springer LNEMS *Proceedings of the Ninth International Workshop on Computer-Aided Scheduling of Public Transport (CASPT)*. To appear.
13. Christian Liebchen and Romeo Rizzi. Cycles bases of graphs. Technical Report 2005-018, TU Berlin, Mathematical Institute, 2005. accepted for publication in Discrete Applied Mathematics.
14. Karl Nachtigall. *Periodic Network Optimization and Fixed Interval Timetables*. Habilitation thesis, Universität Hildesheim, 1998.
15. Michiel A. Odijk. A constraint generation algorithm for the construction of periodic railway timetables. *Transp. Res. B*, 30(6):455–464, 1996.
16. Leon W.P. Peeters. *Cyclic Railway Timetable Optimization*. Ph.D. thesis, Erasmus Universiteit Rotterdam, 2003.
17. Alexander Schrijver and Adri G. Steenbeek. Dienstregelingontwikkeling voor Railned. Rapport CADANS 1.0, Centrum voor Wiskunde en Informatica, December 1994. In Dutch.
18. Paolo Serafini and Walter Ukovich. A mathematical model for periodic scheduling problems. *SIAM Journal on Discrete Mathematics*, 2(4):550–581, 1989.

Determining SMB Superstructures by Mixed-Integer Optimal Control

Sebastian Sager¹, Moritz Diehl², Gundeep Singh³, Achim Küpper⁴, and Sebastian Engell⁴

¹ Interdisciplinary Center for Scientific Computing, Universität Heidelberg
`sebastian.sager@iwr.uni-heidelberg.de`

² Electrical Engineering Department, K.U. Leuven

³ Chemical Engineering Department, IIT Guwahati

⁴ Lehrstuhl für Anlagensteuerungstechnik, Universität Dortmund

1 Introduction

We treat a simplified model of a Simulated Moving Bed (SMB) chromatographic separation process that contains time-dependent discrete decisions. SMB processes have been gaining increased attention lately, see [3], [4] for further references. The related optimization problems are challenging from a mathematical point of view, as they combine periodic nonlinear optimal control problems in partial differential equations (PDE) with time-dependent discrete decisions. For this problem class of mixed-integer optimal control problems (MIOCP) a novel numerical method, developed in [5], is applied.

2 Model

SMB chromatography finds various industrial applications such as sugar, food, petrochemical and pharmaceutical industries. A SMB unit consists of multiple columns filled with solid adsorbent. The columns are connected in a continuous cycle. There are two inlet streams, *desorbent* (De) and *feed* (Fe), and two outlet streams, *raffinate* (Ra) and *extract* (Ex). The continuous countercurrent operation is simulated by switching the four streams periodically in the direction of the liquid flow in the columns, thereby leading to better separation. Due to this discrete switching of columns, SMB processes reach a cyclic or periodic steady state, i.e., the concentration profiles at the end of a period are equal to those at the beginning shifted by one column ahead in direction of the fluid flow. A number of different operating schemes have been proposed to further improve the performance of SMB. These schemes can be described as special cases of the MIOCP (13).

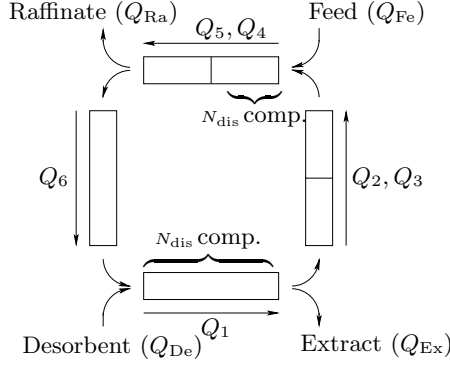


Fig. 1. In standard SMB, the positions of the ports for the four inlet and outlet flows are fixed and constant in time during one period.

The considered SMB unit consists of $N_{\text{col}} = 6$ columns. The flow rate through column i is denoted by Q_i , $i \in I := \{1, \dots, N_{\text{col}}\}$. The raffinate, desorbant, extract and feed flow rates are denoted by Q_{Ra} , Q_{De} , Q_{Ex} and Q_{Fe} , respectively. The (possibly) time-dependent value $w_{i\alpha}(t) \in \{0, 1\}$ denotes if the port of flow $\alpha \in \{\text{Ra}, \text{De}, \text{Ex}, \text{Fe}\}$ is positioned at column $i \in I$. As in many practical realizations of SMB processes only one pump per flow is available and the ports are switched by a 0–1 valve, we obtain the additional *special ordered set type one* restriction

$$\sum_{i \in I} w_{i\alpha}(t) = 1, \quad \forall t \in [0, T], \quad \alpha \in \{\text{Ra}, \text{De}, \text{Ex}, \text{Fe}\}. \quad (1)$$

The flow rates Q_1 , Q_{De} , Q_{Ex} and Q_{Fe} enter as control functions $u(\cdot)$ resp. time-invariant parameters p into the optimization problem, depending on the operating scheme to be optimized. The remaining flow rates are derived by mass balance as

$$Q_{\text{Ra}} = Q_{\text{De}} - Q_{\text{Ex}} + Q_{\text{Fe}} \quad (2)$$

$$Q_i = Q_{i-1} - \sum_{\alpha \in \{\text{Ra}, \text{Ex}\}} w_{i\alpha} Q_{\alpha} + \sum_{\alpha \in \{\text{De}, \text{Fe}\}} w_{i\alpha} Q_{\alpha} \quad (3)$$

for $i = 2, \dots, N_{\text{col}}$. The feed contains two components A and B dissolved in desorbant, with concentrations $c_{\text{Fe}}^{\text{A}} = c_{\text{Fe}}^{\text{B}} = 0.1$. The concentrations of A and B in desorbant are $c_{\text{De}}^{\text{A}} = c_{\text{De}}^{\text{B}} = 0$.

For our case study we use the simplified model presented in [2] for the dynamics in each column that considers axial convection and axial mixing introduced by dividing the respective column into N_{dis} perfectly mixed compartments. Although this simple discretization does not consider all effects present in the advection–diffusion equation for the time and space dependent concentrations, the qualitative behavior of the concentration profiles moving at different velocities through the respective columns is sufficiently well represented for our purposes. We assume that the compartment concentrations are constant. We denote the concentrations of A and B in the compartment with index i by c_i^{A} , c_i^{B} and leave away the time dependency. For the first compartment $j = (i-1)N_{\text{dis}} + 1$ of column $i \in I$ we have by mass transfer for $K = \text{A}, \text{B}$

$$\frac{\dot{c}_j^K}{k^K} = Q_{i^-} c_{j^-}^K - Q_i c_j^K - \sum_{\alpha \in \{\text{Ra}, \text{Ex}\}} w_{i\alpha} Q_\alpha c_{j^-}^K + \sum_{\alpha \in \{\text{De}, \text{Fe}\}} w_{i\alpha} Q_\alpha c_\alpha^K \quad (4)$$

where i^- is the predecesing column, $i^- = N_{\text{col}}$ if $i = 1$, $i^- = i - 1$, else and equivalently $j^- = N$ if $j = 1$, $j^- = j - 1$, else. k^K denotes the axial convection in the column, $k^A = 2N_{\text{dis}}$ and $k^B = N_{\text{dis}}$. Component A is less adsorbed, thus travels faster and is prevailing in the raffinate, while B travels slower and is prevailing in the extract. For interior compartments j in column i we have

$$\frac{\dot{c}_j^K}{k^K} = Q_{i^-} c_{j^-}^K - Q_i c_j^K. \quad (5)$$

The compositions of extract and raffinate, $\alpha \in \{\text{Ex}, \text{Ra}\}$, are given by

$$\dot{M}_\alpha^K = Q_\alpha \sum_{i \in I} w_{i\alpha} c_{j(i)}^K \quad (6)$$

with $j(i)$ the last compartment of column i^- . The feed consumption is

$$\dot{M}_{\text{Fe}} = Q_{\text{Fe}}. \quad (7)$$

These are altogether $2N + 5$ differential equations for the differential states $x = (x_A, x_B, x_M)$ with $x_A = (c_0^A, \dots, c_N^A)$, $x_B = (c_0^B, \dots, c_N^B)$, and finally $x_M = (M_{\text{Ex}}^A, M_{\text{Ex}}^B, M_{\text{Ra}}^A, M_{\text{Ra}}^B, M_{\text{Fe}})$. They can be summarized as

$$\dot{x}(t) = f(x(t), u(t), w(t), p). \quad (8)$$

We define a linear operator $P : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x}$ that shifts the concentration profiles by one column and sets the auxiliary states to zero, i.e.,

$$\begin{aligned} x &\mapsto Px := (P_A x_A, P_B x_B, P_M x_M) \quad \text{with} \\ P_A x_A &:= (c_{N_{\text{dis}}+1}^A, \dots, c_N^A, c_1^A, \dots, c_{N_{\text{dis}}}^A), \\ P_B x_B &:= (c_{N_{\text{dis}}+1}^B, \dots, c_N^B, c_1^B, \dots, c_{N_{\text{dis}}}^B), \\ P_M x_M &:= (0, 0, 0, 0, 0). \end{aligned}$$

Then we can impose periodicity after the unknown cycle duration T by requiring $x(0) = Px(T)$. The purity of component A in the raffinate at the end of the cycle must be higher than $p_{\text{Ra}} = 0.95$ and the purity of B in the extract must be higher than $p_{\text{Ex}} = 0.95$, i.e., we impose the terminal purity conditions

$$M_{\text{Ex}}^A(T) \leq \frac{1 - p_{\text{Ex}}}{p_{\text{Ex}}} M_{\text{Ex}}^B(T), \quad (9)$$

$$M_{\text{Ra}}^B(T) \leq \frac{1 - p_{\text{Ra}}}{p_{\text{Ra}}} M_{\text{Ra}}^A(T). \quad (10)$$

We impose lower and upper bounds on all external and internal flow rates,

$$0 \leq Q_{\text{Ra}}, Q_{\text{De}}, Q_{\text{Ex}}, Q_{\text{Fe}}, Q_1, Q_2, Q_3, Q_4, Q_5, Q_6 \leq Q_{\text{max}} = 2. \quad (11)$$

To avoid draining inflow into outflow streams without going through a column,

$$Q_i - w_{i\text{De}} Q_{\text{De}} - w_{i\text{Fe}} Q_{\text{Fe}} \geq 0 \quad (12)$$

has to hold for all $i \in I$. The objective is to maximize the feed throughput $M_{\text{Fe}}(T)/T$. Summarizing, we obtain the following MIOCP

$$\begin{aligned}
 & \min_{x(\cdot), u(\cdot), w(\cdot), p, T} M_{\text{Fe}}(T)/T \\
 & \text{subject to} \quad \dot{x}(t) = f(x(t), u(t), w(t), p), \\
 & \quad \quad \quad x(0) = Px(T), \\
 & \quad \quad \quad \text{Constraints (9)–(12),} \\
 & \quad \quad \quad \sum_{i \in I} w_{i\alpha}(t) = 1, \quad \forall t \in [0, T], \alpha \in \{\text{Ra, De, Ex, Fe}\}, \\
 & \quad \quad \quad w(t) \in \{0, 1\}^{4N_{\text{col}}}, \quad \forall t \in [0, T].
 \end{aligned} \tag{13}$$

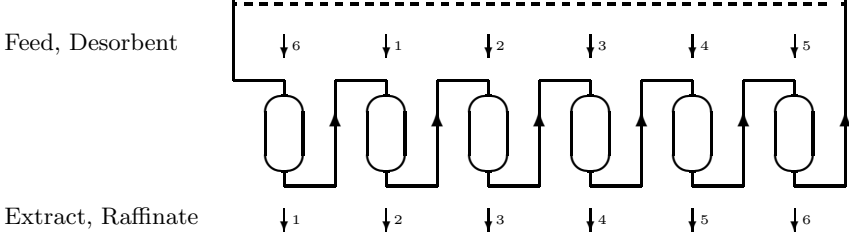
3 Algorithm

Problem (13) fits into the problem class investigated in [5]. We specify a tolerance $\varepsilon > 0$ as problem specific input, determining how large the gap between relaxed and binary solution may be. Furthermore, an initial control discretization grid $\mathcal{G}^0$ is supplied for which an admissible trajectory of the relaxed problem exists. The algorithm MS MINTOC we apply reads as follows:

1. Relax problem (13) to $\tilde{w}(\cdot) \in [0, 1]^{n_{\tilde{w}}}$.
2. Solve this problem for control discretization $\mathcal{G}^0$, obtain the grid-dependent optimal value $\Phi_{\mathcal{G}^0}^{\text{RB}}$ of the trajectory $\mathcal{T}^0$.
3. Refine control discretization grid n_{ext} times as described in [5] and obtain $\Phi^{\text{RB}} = \Phi_{\mathcal{G}^{n_{\text{ext}}}}^{\text{RB}}$ as the objective function value on the finest grid $\mathcal{G}^{n_{\text{ext}}}$.
4. If $\mathcal{T} = \mathcal{T}^{n_{\text{ext}}}$ is binary admissible then STOP, else $k = n_{\text{ext}}$.
5. REPEAT
 - a) Apply a rounding heuristics to $\mathcal{T}$, see [5].
 - b) Use switching time optimization, see [5], initialized with the rounded solution of the previous step. If the obtained trajectory is binary admissible, obtain upper bound Φ^{STO} . If $\Phi^{\text{STO}} < \Phi^{\text{RB}} + \varepsilon$ then STOP.
 - c) Refine the control discretization grid $\mathcal{G}^k$, based on the control values of trajectory $\mathcal{T}$.
 - d) Solve relaxed problem, $\mathcal{T} = \mathcal{T}^k$, $k = k + 1$.

For the solution of subproblems we apply Bock's *direct multiple shooting method* [1] that transforms the optimal control problem into a structured finite-dimensional problem that can be solved with tailored iterative methods, e.g., by sequential quadratic programming (SQP).

4 Numerical Results



Process	Time	Port 1	Port 2	Port 3	Port 4	Port 5	Port 6
SMB fix	0.00 – 0.63	De	Ex		Fe		Ra
SMB relaxed	0.00 – 0.50	De,Ex	Ex		Fe		Ra
PowerFeed	0.00 – 0.56	De	Ex		Fe		Ra
VARICOL	0.00 – 0.18	De	Ex		Fe		Ra
	0.18 – 0.36	De		Ex	Fe		Ra
	0.36 – 0.46	De,Ra		Ex	Fe		
	0.46 – 0.53	De,Ra		Ex		Fe	
Superstruct	0.00 – 0.10	Ex					De
	0.10 – 0.18		De,Ex				
	0.18 – 0.24	De					Ra
	0.24 – 0.49	De		Ex	Fe		Ra
	0.49 – 0.49		De,Ex				

Process	CPU time	Iterations	Objective
SMB fix	10 sec	12	0.7345
SMB relaxed	26 sec	16	0.8747
PowerFeed	290 sec	96	0.8452
VARICOL	298 sec	168	0.9308
Superstruct	3230 sec	290	1.0154

Table 1. Above: fixed or optimized port assignment $w_{i\alpha}$ and switching times of the different process strategies. Below: optimal objective function values, CPU time in seconds and number of SQP iterations needed to find the optimal solution.

We optimized different operation schemes that fit into the general problem formulation (13): **SMB fix**. The $w_{i\alpha}$ are fixed as shown in the table. The flow rates Q are constant in time, i.e., they enter as optimization parameters p into (13). **SMB relaxed**. As above. But the $w_{i\alpha}$ are free for optimization and relaxed to $w_{i\alpha} \in [0,1]$, allowing for a "splitting" of the ports. In **PowerFeed** the flow rates are modulated during one period, i.e., the Q enter as control functions $u(\cdot)$ into (13). **VARICOL**. The ports switch asynchronously, but in a given order as seen in the table. The switching times are subject to optimization. **Superstruct**. This scheme is the

most general and allows for arbitrary switching of the ports. The flow rates enter as continuous control functions, but are found to be bang-bang by the optimizer (i.e., whenever the port is given in the table, the respective flow rate is at its upper bound).

All calculations were done on a Pentium 1.7 GHz desktop computer, using the optimal control software package MUSCOD-II.

5 Conclusions and Outlook

We have presented a benchmark problem for mixed-integer optimal control methods that is an extension of a recently published benchmark problem without integer variables. This problem shows the same qualitative behavior as more rigorous first-principle models of simulated moving bed processes and is therefore well suited for a comparison of different process strategies.

We applied a novel algorithm to treat the occurring time-dependent integer decisions. It deals with the integrality by a combination of adaptivity and tailored heuristics, while theoretical results guarantee a user-defined integrality gap. It is very efficient as continuous subproblems are solved in a homotopy, especially if the solution of the relaxed problem is bang-bang on a suitable discretization grid, as is the case for the problem under consideration here. This result is an extension of the findings of [4]. Here, it was observed that for the determination of a *time-independent* optimal zone configuration a relaxation of the binary variables sufficed, as the optimal relaxed solution was integral. However, as seen in the case of "SMB relaxed", this must not always be the case, especially if different objective functions are chosen or additional constraints are involved that lead to path-constrained arcs of the optimal solution. Further research should concentrate on the question, when this is the case — the proposed algorithm will also work for these problems.

References

1. H.G. Bock and K.J. Plitt. A multiple shooting algorithm for direct solution of optimal control problems. In *Proceedings 9th IFAC World Congress Budapest*, pages 243–247. Pergamon Press, 1984.
2. M. Diehl and A. Walther. A test problem for periodic optimal control algorithms. Technical report, ESAT/SISTA, K.U. Leuven, 2006.
3. S. Engell and A. Toumi. Optimisation and control of chromatography. *Computers and Chemical Engineering*, 29:1243–1252, 2005.
4. Y. Kawajiri and L.T. Biegler. Large-scale optimization strategies for zone configuration of simulated moving beds. In *16th European Symposium on Computer Aided Process Engineering and 9th International Symposium on Process Systems Engineering*. Elsevier, 2006.
5. S. Sager. *Numerical methods for mixed-integer optimal control problems*. Der andere Verlag, Tönning, Lübeck, Marburg, 2005. ISBN 3-89959-416-9. Available at <http://sager1.de/sebastian/downloads/Sager2005.pdf>.

GOR Diplomarbeitspreis 2006

Complexity of Pure-Strategy Nash Equilibria in Non-Cooperative Games

Juliane Dunkel

Operations Research Center, Massachusetts Institute of Technology, U.S.A.
juliane@mit.edu

1 Introduction

Game theory in general and the concept of Nash equilibrium in particular have lately come under increased scrutiny by theoretical computer scientists. Computing a mixed Nash equilibrium is a case in point. For many years, one of the most important open problems was the complexity of computing a mixed Nash equilibrium in games with only two players. Only recently was it solved by a sequence of significant papers (Goldberg and Papadimitriou (2006), Daskalakis et.al. (2006), Chen and Deng (2005), Daskalakis and Papadimitriou (2005), and Chen and Deng (2006)).

While Nash (1951) showed that mixed Nash equilibria do exist in any finite noncooperative game, it is well known that pure-strategy Nash equilibria are in general not guaranteed to exist. It is therefore natural to ask which games have pure-strategy Nash equilibria and, if applicable, how difficult is it to find one. In this article, we study these questions for certain classes of weighted congestion and local-effect games.

Congestion games are a fundamental class of noncooperative games possessing pure-strategy Nash equilibria. Their wide applicability (e.g., in network routing and production planning) has made them an object of extensive study. In the network version, each player wants to route one unit of flow on a path from her origin to her destination at minimum cost, and the cost of using an arc only depends on the total number of players using that arc. A natural extension is to allow for players sending different amounts of flow, which results in so-called weighted congestion games. While examples have been exhibited showing that pure-strategy Nash equilibria need not exist, we prove that it actually is NP-hard to determine whether a given weighted network congestion game has a pure-strategy Nash equilibrium. This is true regardless of whether flow is unsplittable (has to be routed on a single path for each player) or not.

A related family of games are local-effect games, where the disutility of a player taking a particular action depends on the number of players taking the same action and on the number of players choosing related actions. We show that the problem of deciding whether a bidirectional local-effect game has a pure-strategy Nash equilibrium is NP-complete, and that the problem of finding a pure-strategy Nash

equilibrium in a bidirectional local-effect game with linear local-effect functions (for which the existence of a pure-strategy Nash equilibrium is guaranteed) is PLS-complete. The latter proof uses a tight PLS-reduction, which implies the existence of instances and initial states for which any sequence of selfish improvement steps needs exponential time to reach a pure-strategy Nash equilibrium.

Due to space limitations, proofs are only sketched or omitted completely. Details can be found in Dunkel (2005).

2 Weighted Congestion Games

In an *unweighted congestion game*, we are given a set of players $N = \{1, 2, \dots, n\}$, and a set of resources E . For each player $i \in N$, her set S_i of available strategies is a collection of subsets of the resources. A cost function $f_e : \mathbb{N} \rightarrow \mathbb{R}_+$ is associated with each resource $e \in E$. Given a strategy profile $s = (s_1, s_2, \dots, s_n) \in S = S_1 \times S_2 \times \dots \times S_n$, the cost of player i is $c_i(s) = \sum_{e \in s_i} f_e(n_e(s))$, where $n_e(s)$ denotes the number of players using resource e in s . In other words, in a congestion game each player chooses a subset of resources that are available to her, and the cost to a player is the sum of the costs of the resources used by her, where the cost of a resource only depends on the total number of players using this resource.

In a *network congestion game*, the arcs of an underlying directed network represent the resources. Each player $i \in N$ has an origin-destination pair (a_i, b_i) of nodes, and the set S_i of pure strategies available to player i is the set of directed (simple) paths from a_i to b_i . In a symmetric network congestion game all players have the same origin-destination pair.

In a *weighted network congestion game*, each player $i \in N$ has a positive integer weight w_i , which constitutes the amount of flow she wants to ship from a_i to b_i . In the case of unsplittable flows, the cost of player i adopting strategy s_i in a strategy profile $s = (s_1, s_2, \dots, s_n) \in S$ is given by $c_i(s) = \sum_{e \in s_i} f_e(\theta_e(s))$, where $\theta_e(s) = \sum_{i: e \in s_i} w_i$ denotes the total flow on arc e in s . In integer-splittable network congestion games, a player with weight greater than one can choose a subset of paths on which to route her flow simultaneously.

While every unweighted congestion game possesses a pure-strategy Nash equilibrium (Rosenthal 1973), this is not true for weighted congestion games; see, e.g., Fig. 1 in Fotakis, Kontogiannis, and Spirakis (2005). We can actually turn their instance into a gadget to derive the following result.

Theorem 1. *The problem of deciding whether a weighted symmetric network congestion game with unsplittable flows possesses a pure-strategy Nash equilibrium is NP-complete.*

For network congestion games with integer-splittable flows, we obtain the following result.

Theorem 2. *The problem of deciding whether a weighted network congestion game with integer-splittable flows possesses a pure-strategy Nash equilibrium is strongly NP-hard. Hardness even holds if there is only one player with weight 2, and all other players have unit weights.*

Proof. Consider an instance of MONOTONE3SAT with set of variables $X = \{x_1, x_2, \dots, x_n\}$ and set of clauses $C = \{c_1, c_2, \dots, c_m\}$. We construct a game that has one player p_x for every variable $x \in X$ with weight $w_x = 1$, origin x and destination $\bar{x}$. Moreover, each clause $c \in C$ gives rise to a player p_c with weight $w_c = 1$, origin c , and destination $\bar{c}$. There are three more players p_1, p_2 , and p_3 with weights $w_1 = 1, w_2 = 2, w_3 = 1$ and origin-destination pairs $(s, t_1), (s, t_2), (s, t_3)$, respectively. For every variable $x \in X$ there are two disjoint paths P_x^1, P_x^0 from x to $\bar{x}$ in the network. Path P_x^0 consists of $2|\{c \in C | x \in c\}| + 1$ arcs and P_x^1 has $2|\{c \in C | \bar{x} \in c\}| + 1$ arcs with cost functions as shown in Fig. 1. For each pair $(c, \bar{c})$, we construct two disjoint paths P_c^1, P_c^0 from c to $\bar{c}$. Path P_c^1 consists of only two arcs. The paths P_c^0 will have seven arcs each and are constructed for $c = c_j$ in the order $j = 1, 2, \dots, m$ as follows. For a positive clause $c = c_j = (x_{j_1} \vee x_{j_2} \vee x_{j_3})$ with $j_1 < j_2 < j_3$, path P_c^0 starts with the arc connecting c to the first inner node v_1 on path $P_{x_{j_1}}^1$ that has only two incident arcs so far. The second arc is the unique arc (v_1, v_2) of path $P_{x_{j_1}}^1$ that has v_1 as its start vertex. The third arc connects v_2 to the first inner node v_3 on path $P_{x_{j_2}}^1$ that has only two incident arcs so far. The fourth arc is the only arc (v_3, v_4) on $P_{x_{j_2}}^1$ with start vertex v_3 . From v_4 , there is an arc to the first inner node v_5 on $P_{x_{j_3}}^1$ that has only two incident arcs so far, followed by (v_5, v_6) of $P_{x_{j_3}}^1$. The last arc of P_c^0 connects v_6 to $\bar{c}$ (see Fig. 1). For a negative clause $c = c_j = (\bar{x}_{j_1} \vee \bar{x}_{j_2} \vee \bar{x}_{j_3})$ we proceed in the same way, except that we choose the inner nodes v_i from the upper variable paths $P_{x_{j_1}}^0, P_{x_{j_2}}^0, P_{x_{j_3}}^0$. The

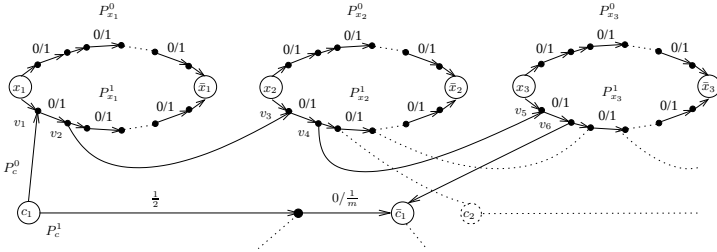


Fig. 1. Part of the constructed network corresponding to a positive clause $c_1 = (x_1 \vee x_2 \vee x_3)$. The notation a/b defines a cost per unit flow of value a for load 1 and b for load 2. Arcs without specified values have zero cost.

strategy set of player p_x is $\{P_x^1, P_x^0\}$. We will interpret it as setting the variable x to *true* (*false*) if p_x sends her unit of flow over P_x^1 (P_x^0). Note that player p_c can only choose between the paths P_c^1 and P_c^0 . This is due to the order in which the paths $P_{c_j}^0$ are constructed. Depending on whether player p_c sends her unit of flow over path P_c^1 or P_c^0 , the clause c will be satisfied or not.

The second part of the network consists of all origin-destination pairs and paths for players p_1, p_2, p_3 (see Fig. 2). Player p_1 can choose between paths $U_1 = \{(s, t_2), (t_2, t_1)\}$ and $L_1 = \{(s, t_1)\}$. Player p_2 is the only player who can split her flow; that is, she can route her two units either both over path $U_2 = \{(s, t_2)\}$, both over path $L_2 = \{(s, t_1), (t_1, t_2)\}$, or one unit on the upper and the other unit

on the lower path; i.e., $S_2 = \{L_2, U_2, LU_2\}$. Finally, player p_3 has three possible paths to choose from, an upper path U_3 and paths $M_3 = \{(s, t_2), (t_2, t_3)\}$ and $L_3 = \{(s, t_1), (t_1, t_2), (t_2, t_3)\}$ (see Fig. 2).

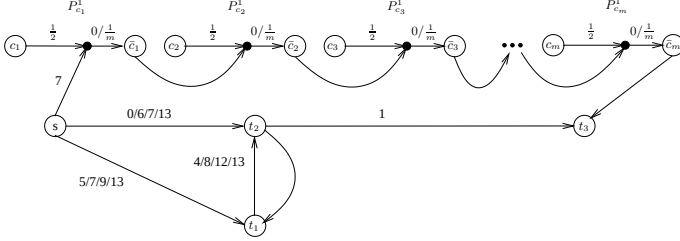


Fig. 2. Part of the constructed network that is used by players p_1 , p_2 , and p_3 . A single number a on an arc defines a constant cost a per unit flow for this arc.

Given a satisfying truth assignment, we define a strategy state s of the game by setting the strategy of player p_x to be P_x^1 for a true variable x , and P_x^0 otherwise. Each player p_c plays P_c^1 . Furthermore, $s_1 = L_1$, $s_2 = U_2$, and $s_3 = M_3$. It is easy to show that s is a pure-strategy Nash equilibrium.

For the other direction, we first observe that any pure-strategy Nash equilibrium s of the game has the following properties: (a) player p_3 does not use path U_3 , (b) for the cost of player p_3 we have $c_3(s) \geq 8$, and (c) each player p_c routes her unit flow over path P_c^1 . Property (a) follows from the fact that the sub-game induced by the nodes s, t_1, t_2 and players p_1 and p_2 only does not have a pure-strategy Nash equilibrium. Property (a) implies (b), and property (c) can be proved by contradiction assuming (a) and (b). Using the properties of the Nash equilibrium s , it is now easy to show that the truth assignment that sets a variable x to *true* if the corresponding player uses P_x^1 , and *false* otherwise, satisfies all clauses. $\square$

3 Local-Effect Games

In a *local-effect game*, we are given a set of players $N = \{1, 2, \dots, n\}$ and a common action set A available to each player. For each pair of actions $a, a' \in A$, a cost function $F_{a',a} : \mathbb{Z}_+ \rightarrow \mathbb{R}_+$ expresses the impact of action a' on the cost of action a , which depends only on the number of players that choose action a' . For $a, a' \in A$ with $a \neq a'$, $F_{a',a}$ is called a *local-effect function*, and it is assumed that $F_{a',a}(0) = 0$. Moreover, the local-effect function $F_{a',a}$ is either strictly increasing or identical zero. For a given strategy state $s = (s_1, s_2, \dots, s_n) \in A^n$, n_a denotes the number of players playing action a in s . The cost to a player $i \in N$ for playing action s_i in strategy state s is given by $c_i(s) = F_{s_i, s_i}(n_{s_i}) + \sum_{a \in A, a \neq s_i} F_{a, s_i}(n_a)$. A local-effect game is called a *bidirectional* local-effect game if for all $a, a' \in A$, $a \neq a'$, and for all $x \in \mathbb{Z}_+$, $F_{a',a}(x) = F_{a,a'}(x)$.

Leyton-Brown and Tennenholtz (2003) gave a characterization of local-effect games that have an exact potential function and which are therefore guaranteed to

possess pure-strategy Nash equilibria. One of these subclasses are bidirectional local-effect games with linear local-effect functions. However, without linear local-effect functions, deciding the existence is hard.

3.1 Computational Complexity

Theorem 3. *The problem of deciding whether a bidirectional local-effect game has a pure-strategy Nash equilibrium is NP-complete.*

The next result implies that computing a pure-strategy Nash equilibrium for a bidirectional local-effect game with linear local-effect functions is as least as hard as finding a local optimum for several combinatorial optimization problems with efficiently searchable neighborhoods.

Theorem 4. *The problem of computing a pure-strategy Nash equilibrium for a bidirectional local-effect game with linear local-effect functions is PLS-complete.*

Proof. We reduce from POSNAE3FLIP (Schäffer and Yannakakis 1991): Given not-all-equal clauses with at most three literals, (x_1, x_2, x_3) or (x_1, x_2) , where x_i is either a variable or a constant (0 or 1), and a weight for each clause, find a truth assignment such that the total weight of satisfied clauses cannot be improved by flipping a single variable.

For simplicity, we assume that we are given an instance of POSNAE3FLIP with set $C = C_2 \dot{\cup} C_3$ of clauses containing two or three variables but no constants, a positive integer weight w_c for each clause $c \in C$, and set of variables $\{x_1, \dots, x_n\}$. We construct a bidirectional local-effect game with linear local-effect functions as follows: There are n players with common action set A that contains two actions a_i and $\bar{a}_i$ for each variable x_i , $i = 1, 2, \dots, n$. Let $M = 2n \sum_{c \in C} w_c + 1$. For each action $a \in A$, $F_{a,a}(x) = 0$ if $x \leq 1$, and $F_{a,a}(x) = M$ otherwise. If $C_i = \{c \in C | x_i \in c\}$ denotes the subset of clauses containing variable x_i , the local-effect functions are given for $i, j \in \{1, 2, \dots, n\}$, $i \neq j$, by

$$F_{a_i, a_j}(x) = F_{\bar{a}_i, \bar{a}_j}(x) = \left(2 \sum_{c \in C_2 \cap C_i \cap C_j} w_c + \sum_{c \in C_3 \cap C_i \cap C_j} w_c \right) x.$$

However, the local-effect functions F_{a_i, a_j} and $F_{\bar{a}_i, \bar{a}_j}$ are zero if there is no clause containing both x_i and x_j . Furthermore, $F_{a_i, \bar{a}_i}(x) = F_{\bar{a}_i, a_i}(x) = Mx$ for all $i \in \{1, 2, \dots, n\}$. All local-effect functions not defined so far are identical zero.

For any solution $s = (s_1, s_2, \dots, s_n)$, $s_i \in A$, of the game, we define the corresponding truth assignment to the variables x_i of the POSNAE3FLIP instance by $x_i = 1$ if $|\{j | s_j = a_i\}| \geq 1$, and $x_i = 0$ otherwise. One can show that any pure-strategy Nash equilibrium s of the game fulfills $|\{j | s_j = a_i\}| + |\{j | s_j = \bar{a}_i\}| = 1$; and the corresponding truth assignment is indeed a local optimum of the POSNAE3FLIP instance. $\square$

Since the reduction actually is a *tight* PLS-reduction, we obtain the following results.

Corollary 1. (i) *There are instances of bidirectional local-effect games with linear local-effect functions that have exponentially long shortest improvement paths.* (ii) *For a bidirectional local-effect game with linear local-effect functions, the problem of finding a pure-strategy Nash equilibrium that is reachable from a given strategy state via selfish improvement steps is PSPACE-complete.*

The following result underlines that finding a pure Nash equilibrium for bidirectional local-effect games with linear local-effect functions is indeed hard.

Theorem 5. *Given an instance of a bidirectional local-effect games with linear local-effect functions, a pure-strategy profile s_0 , and an integer $k > 0$ (unarily encoded), it is NP-complete to decide whether there exists a sequence of at most k selfish steps that transforms s_0 to a pure-strategy Nash equilibrium.*

3.2 Pure Price of Stability for Bidirectional Local-Effect Games

We derive bounds on the pure-price of stability for games with linear local-effect functions where the social objective is the sum of the costs of all players.

Theorem 6. *The pure price of stability for bidirectional local-effect games with only linear cost functions is bounded by 2.*

The proof is based on a technique suggested by Anshelevich et al. (2004) using the potential function. By the same technique, we can derive the following bound for the case of quadratic cost-functions and linear local-effect functions.

Theorem 7. *The pure price of stability for bidirectional local-effect games with $F_{a,a}(x) = m_a x^2 + q_a x$, $q_a \geq 0$ for all $a \in A$ and linear local-effect functions is bounded by 3.*

References

1. Anshelevich, E., A. Dasgupta, J. Kleinberg, É. Tardos, T. Wexler, and T. Roughgarden (2004). The price of stability for network design with fair cost allocation. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, Rome, Italy, pp. 295–304.
2. Chen, X. and X. Deng (2005). 3-Nash is PPAD-complete. *Electronic Colloquium on Computational Complexity* TR05-134.
3. Chen, X. and X. Deng (2006). Settling the complexity of 2-player Nash-equilibrium. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, Berkeley, CA, to appear.
4. Daskalakis, C., P. Goldberg, and C. Papadimitriou (2006). The complexity of computing a Nash equilibrium. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, Seattle, WA, pp. 71–78.
5. Daskalakis, C. and C. Papadimitriou (2005). Three-player games are hard. *Electronic Colloquium on Computational Complexity* TR05-139.
6. Dunkel, J. (2005). The Complexity of Pure-Strategy Nash Equilibria in Non-Cooperative Games. Diplomarbeit, Institute of Mathematics, Technische Universität Berlin, Germany, July 2005.

7. Fotakis, D., S. Kontogiannis, and P. Spirakis (2005). Selfish unsplittable flows. *Theoretical Computer Science* 348, pp. 226–239.
8. Goldberg, P. and C. Papadimitriou (2006). Reducibility among equilibrium problems. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, Seattle, WA, pp. 61–70.
9. Leyton-Brown, K. and M. Tennenholtz (2003). Local-effect games. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, pp. 772–780.
10. Nash, J. (1951). Non-cooperative games. *Annals of Mathematics* 54, 268–295.
11. Rosenthal, R. (1973). A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory* 2, pp. 65–67.
12. Schäffer, A. and M. Yannakakis (1991). Simple local search problems that are hard to solve. *SIAM Journal on Computing* 20, pp. 56–87.

Traffic Optimization Under Route Constraints with Lagrangian Relaxation and Cutting Plane Methods

Felix G. König

Fakultät II - Mathematik und Naturwissenschaften, Technische Universität Berlin, Germany
fkoenig@math.tu-berlin.de

Summary. The optimization of traffic flow in congested urban road networks faces a well-known dilemma: Optimizing system performance is unfair with respect to the individual drivers' travel times; and a fair user equilibrium may result in bad system performance. As a remedy, computing a system optimum with fairness conditions, realized by length constraints on the routes actually used by drivers, has been suggested in [5]. This poses interesting mathematical challenges, namely the non-linearity of the objective function and the necessity to deal with path constraints in large networks. While the authors present results suggesting that solutions to this constrained system optimum problem (CSO) are indeed equally good and fair, they rely on a standard Frank-Wolfe/Partan-algorithm to obtain them.

In this paper, we present a Lagrangian relaxation of the CSO problem for which the Lagrangian dual function can be evaluated by a decomposition into constrained shortest path problems which we solve exactly employing state-of-the-art acceleration techniques. The Lagrangian dual problem is then solved by a special cutting plane method.

Finally, we obtain test results which suggest that this approach outperforms previously described solution schemes for the CSO problem.

1 The Constrained System Optimum Problem

The street network is represented by a digraph $G = (V, A)$ where each arc is equipped with an increasing differentiable latency function $l_a : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ associating a latency time $l_a(x_a)$ with a given flow rate x_a on arc a . We use the following latency function with free flow travel time l_a° and tuning parameters α and β suggested by the U. S. Bureau of Public Roads:

$$l_a(x_a) := l_a^\circ \left(1 + \alpha \left(\frac{x_a}{c_a} \right)^\beta \right).$$

Additionally, we define a “normal length” τ_a for each arc which will be used to control the fairness of the calculated route suggestions. Technically, τ_a may be any

metric on A ; in order to achieve equally good and fair routes, we follow the suggestion of the authors of [5] and choose τ_a to be the latency incurred by drivers on arc a in a user equilibrium.

The set K of traffic demands is represented by origin destination pairs $(s_k, t_k) \in V \times V$ with an associated demand d_k for each $k \in K$. Let P_k represent the set of all possible paths for commodity k . Let $P := \cup_{k \in K} P_k$ denote the set of all paths in G possibly used and $\tau_p := \sum_{a \in p} \tau_a$ the normal travel time incurred on $p \in P$.

The parameter $\varphi \geq 1$ defines the feasibility of routes: a path $p \in P_k$ is feasible iff $\tau_p \leq \varphi T_k$, $T_k := \min_{p \in P_k} \tau_p$ being the length of a shortest s_k - t_k path according to normal length. In a certain sense, φ represents the factor by which a driver's travel time may increase in our solution compared to the user equilibrium.

With z_a^k representing the flow rate of commodity k on arc a and x_p the flow rate on path p , the CSO problem may be stated as the following non-linear multi-commodity flow problem with path constraints:

$$\text{Minimize} \quad \sum_{a \in A} l_a(x_a) x_a \quad (1a)$$

$$\text{subject to} \quad \sum_{k \in K} z_a^k = x_a \quad a \in A \quad (1b)$$

$$\sum_{p \in P_k: a \in p} x_p = z_a^k \quad a \in A; k \in K \quad (1c)$$

$$\sum_{p \in P_k} x_p = d_k \quad k \in K \quad (1d)$$

$$\tau_p \leq \varphi T_k \quad p \in P_k : x_p > 0; k \in K \quad (1e)$$

$$x_p \geq 0 \quad p \in P.$$

Note that the objective (1a) is convex; furthermore, the convex set of feasible solutions may be artificially bounded by $x_a \leq D_k := \sum_{k \in K} d_k$ and thus be made compact.

2 Lagrangian Relaxation for the CSO Problem

We now state a Lagrangian relaxation of (1) in which we drop constraints (1b) coupling the flow rates for each commodity z_a^k with the total arc flows x_a . A similar idea is used in [2] to solve different variants of the multi-commodity flow problem. With Lagrangian multipliers u_a , the resulting relaxation for the CSO problem is

$$\text{Minimize} \quad L(x, z, u) := \sum_{a \in A} \left(l_a(x_a) x_a + u_a \left(\sum_{k \in K} z_a^k - x_a \right) \right) \quad (2a)$$

$$\text{subject to} \quad \sum_{p \in P_k: a \in p} x_p = z_a^k \quad a \in A; \quad k \in K \quad (2b)$$

$$\sum_{p \in P_k} x_p = d_k \quad k \in K \quad (2c)$$

$$\tau_p \leq \varphi T_k \quad p \in P_k : x_p > 0; \quad k \in K \quad (2d)$$

$$x_p \geq 0 \quad p \in P.$$

By separating this problem by variables z and x , we can restate it as two independent minimization problems:

$$\text{Minimize} \quad L_1(z, u) := \sum_{k \in K} \left(\sum_{a \in A} u_a z_a^k \right) \quad (3a)$$

$$\text{subject to} \quad \sum_{p \in P_k: a \in p} x_p = z_a^k \quad a \in A; \quad k \in K \quad (3b)$$

$$\sum_{p \in P_k} x_p = d_k \quad k \in K \quad (3c)$$

$$\tau_p \leq \varphi T_k \quad p \in P_k : x_p > 0; \quad k \in K \quad (3d)$$

$$x_p \geq 0 \quad p \in P$$

and

$$\text{Minimize} \quad L_2(x, u) := \sum_{a \in A} (l_a(x_a) x_a - u_a x_a) \quad (4a)$$

$$\text{subject to} \quad x_a \geq 0 \quad a \in A.$$

Problem (3) corresponds to solving $|K|$ standard constrained shortest path problems CSP_k , $k \in K$:

$$CSP_k : \text{Find an } s_k\text{-}t_k\text{-path } p_k^* \text{ with } \tau_{p_k^*} \leq \varphi T_k \text{ and } \sum_{a \in p_k^*} u_a \text{ minimal.} \quad (5)$$

For given u and optimal solutions p_k^* to problems (5), let

$$L_1^*(u) := \sum_{k \in K} \left(d_k \cdot \sum_{a \in p_k^*} u_a \right) \quad (6)$$

denote the value of an optimal solution to (3). The constrained shortest path problem is (weakly) NP-hard by an obvious reduction from Knapsack; yet algorithms to compute exact solutions to real-world instances efficiently are described in [6]. We use one of the approaches mentioned therein, namely the goal-oriented modified Dijkstra algorithm, to solve problem (5).

Problem (4) consists of $|A|$ simple analytic minimization problems: For given u and for each $a \in A$, we are looking for the minimum of the convex function

$$L_2^a(x, u) := (l_a(x_a) - u_a) \cdot x_a = \left(l_a^\circ \left(1 + \alpha \left(\frac{x_a}{c_a} \right)^\beta \right) - u_a \right) \cdot x_a.$$

The desired minimum exists and is unique as long as l_a is not a constant function. However, for arcs with constant latency, the relaxed constraints (1b) are redundant; we merely replace the corresponding terms in the objective function (1a) of the original problem by

$$\sum_{k \in K} l_a^\circ \cdot z_a^k = l_a^\circ \cdot \sum_{k \in K} z_a^k = l_a^\circ \cdot x_a = l_a(x_a)$$

and thus eliminate constraints (1b) for these arcs.

For all other arcs, the minimum of $L_2^a(x, u)$ is given by

$$L_2^{a*}(u_a) = \begin{cases} c_a \cdot (l_a^\circ - u_a) \cdot \frac{\beta}{\beta+1} \cdot \sqrt[\beta]{\frac{u_a - l_a^\circ}{l_a^\circ \cdot \alpha \cdot (\beta+1)}} & \text{if } l_a^\circ < u_a \\ 0 & \text{else} \end{cases} \quad (7)$$

and the optimal value of (4) by

$$L_2^*(u) = \sum_{a \in A} L_2^{a*}(u_a). \quad (8)$$

For given u , let

$$\theta(u) := L_1^*(u) + L_2^*(u).$$

denote the optimal value of (2). The resulting Lagrangian dual problem may then be stated as

$$\max_{u \in \mathbb{R}^{|A|}} \theta(u). \quad (9)$$

3 Proximal ACCPM to Solve the Dual Lagrangian

In order to solve the Lagrangian dual problem (9), we employ the Analytic Center Cutting Plane Method (ACCPM), first described by Goffin, Haurie and Vial in [4] and later refined by du Merle and Vial in [3]. The Logistics Laboratory at the University of Geneva maintains an implementation of the method which has been successfully applied to a number of problems (see [1] for an overview). We use a variant of the method described in [2], which solves problems of the form

$$\max\{\theta(u) = f_1(u) + f_2(u); u \in \mathbb{R}^n\}, \quad (10)$$

where f_1 is concave and f_2 is self-concordant and concave. The additive components (6) and (8) of our dual objective function (2a) fulfill just that.

In order for ACCPM to function, certain oracles for functions f_1 and f_2 must be provided: Given a point u^t , the first oracle must return the corresponding value $f_1^t := f_1(u^t)$ and a subgradient ξ^t of f_1 at u^t . Similarly, the second oracle must return the value $f_2^t = f_2(u^t)$ and f_2 's first and second derivative at u^t .

In our case, equations (6) and (8) deliver the required function values. Furthermore, the arc flow vector x^t resulting from the sum of all constrained shortest path calculations for u^t delivers the needed subgradient:

$$\begin{aligned} L_1^*(u) &= \min_{x \geq 0} L_1(u, x) \leq L_1(u, x^t) = u^\top x^t \\ &= u^\top x^t + u^{t\top} x^t - u^{t\top} x^t = L_1^*(u^t) + x^{t\top} (u - u^t). \end{aligned}$$

Finally, the first and second derivatives of (8) can easily be computed analytically.

For $t = 1, 2, \dots, T$, let u^t be the points at which the oracles have been queried and $\underline{\theta}$ the best objective value achieved so far. ACCPM now maintains the localization set

$$\begin{aligned} \Omega^T &:= \{(u, z + \zeta) \in \mathbb{R}^n \times \mathbb{R} : \\ &z \leq f_1(u^t) + \xi^{t\top} (u - u^t) \ \forall t = 1, 2, \dots, T; \ \zeta \leq f_2(u); \ z + \zeta \geq \underline{\theta}\}. \end{aligned}$$

which obviously contains all optimal solutions to (10). It is bounded from below by the best objective value and from above by the gradients of the objective function obtained so far. In each iteration, a new query point is selected from the localization set; the oracles then return the corresponding objective value and a new gradient, further bounding the localization set.

The uniqueness of this method lies in its choice of points at which to query the oracles: A query point is chosen to be the proximal analytic center of the current localization set, namely the unique minimizer of a certain barrier function measuring the weighted distances of a point to all planes bounding the localization set, and to the point at which the best objective value so far has been achieved. In [2], a Newton-method to find these minimizers is described and a detailed description of proximal analytic centers and their barrier functions is given.

While $\underline{\theta}$ serves as a trivial lower bound for the dual problem, a feasible primal solution (and thus an upper bound for the dual problem) can be obtained exploiting information generated during the calculation of the proximal analytic centers. The method terminates and returns this feasible primal solution when the distance between the two bounds is less than the user-specified optimality gap.

4 Results

Computational tests were run on instances also used in [5]. Except for **Neukoelln**, all instances are part of the online database of “Transportation Network Problems” (<http://www.bgu.ac.il/~bargera/tntp/>). Table 1 states the instances’ most important properties.

As in [5], the desired optimality gap for all experiments was set to 0.005 (0.5%) and the fairness parameter φ was set to 1.02 to make results as comparable as possible. Because the machine used to conduct the experiments in [5] was unavailable to us, we compare results based on number of iterations; since both the Partan algorithm and our approach rely on constrained shortest path calculations for each commodity in each iteration, this seems reasonable enough.

Figure 1 illustrates how our approach outperforms the Partan algorithm on all instances. The difference is most significant on instance **Sioux Falls**, where the Partan algorithm needs more than six times the iterations needed by ACCPM.

Table 1. Instances used and their most important properties.

Instance	$ V $	$ A $	$ K $
Sioux Falls	24	76	528
Winnipeg	1052	2836	4344
Neukoelln	1890	4040	3166
Chicago Sketch	933	2950	83113

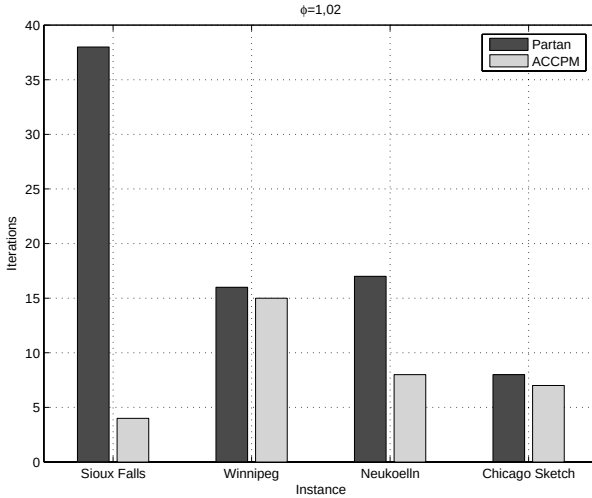


Fig. 1. Iterations needed with Partan/ACCPM to obtain an optimality gap of ≤ 0.005 .

Finally, it is worth mentioning that for every instance tested, more than 97% of the runtime were spent finding constrained shortest paths; the runtime of ACCPM is negligible in our context. This fact also illustrates the importance of a fast algorithm to find constrained shortest paths. Runtimes for all instances almost doubled when a basic labeling algorithm without goal-orientation was used - despite the fact that the goal-oriented approach requires the calculation of a shortest-path-tree for each commodity and iteration as a substantial preprocessing step.

References

1. F. Babonneau, C. Beltran, A. Haurie, C. Tadonki, and J.-P. Vial. Proximal-ACCPM: A versatile oracle based optimization method. *Computational Management Science*, to appear.
2. F. Babonneau and J.-P. Vial. ACCPM with a nonlinear constraint and an active set strategy to solve nonlinear multicommodity flow problems. *Mathematical Programming*, to appear.

3. O. du Merle and J.-P. Vial. Proximal-ACCPM, a cutting plane method for column generation and lagrangian relaxation: Application to the p -median problem. Technical Report 2002.23, Logilab, University of Geneva, 2002.
4. J.-L. Goffin, A. Haurie, and J.-P. Vial. Decomposition and nondifferentiable optimization with the projective algorithm. *Management Science*, 38(2):284–302, 1992.
5. O. Jahn, R. H. Möhring, A. S. Schulz, and N. E. Stier-Moses. System-optimal routing of traffic flows with user constraints in networks with congestion. *Operations Research*, 53:600–616, 2005.
6. E. Köhler, R. H. Möhring, and H. Schilling. Acceleration of shortest path and constrained shortest path computation. In *Proceedings of the WEA 2005*, pages 126–138, 2005.

Fare Planning for Public Transport

Marika Neumann

Konrad-Zuse-Zentrum für Informationstechnik Berlin, Germany
`marika.neumann@zib.de`

Summary. In this paper we investigate the *fare planning model* for public transport, which consists in designing a system of fares maximizing the revenue. We discuss a discrete choice model in which passengers choose between different travel alternatives to express the demand as a function of fares. Furthermore, we give a computational example for the city of Potsdam and discuss some theoretical aspects.

1 Introduction

The design and the level of fares influence the passenger volume and consequently the revenue of a public transport system. Therefore, they are an important instrument to improve the profitability of the public transport system or to achieve other goals, e.g., to provide access to public transport for the general public.

Some articles in the literature deal with different approaches to find optimal fares for public transport. Hamacher and Schöbel [6] develop a model for designing fares and zones maximizing the similarity to a given fare system, e.g., a distance dependent one. Kocur and Hendrickson [7] and De Borger, Mayeres, Proost, and Wouters [5] introduce models for maximizing the revenue and the social welfare, respectively, subject to several budget constraints. The majority of the literature on public transport fares, however, discusses only theoretical concepts, e.g. marginal cost pricing (Pedersen [8]) and price elasticities (Curtin [4]).

In this article, we want to investigate a model to compute the fares that optimize the revenue for the public transport. This model is called the *fare planning model*. The main advantage of our approach is the inclusion of the public transport network. This allow us to distinguish different travel routes, e.g. between means of transportation like bus or subway, between slow and fast, short and long routes. Therefore it is possible to design complex and optimal fare systems.

This work is a summary of the master thesis “Mathematische Preisplanung im ÖPNV”, that I wrote at the Zuse Institute Berlin. The advisor was Prof. Dr. Grötschel. Some parts of the master thesis were published in the Operations Research Proceedings 2005 [3], another summary can be found in [2]. The theoretical results in Section 3.2 include some new aspects.

2 General Fare Planning Model

We consider a traffic network whereas the nodes V represent the stations and $D \subseteq V \times V$ is a set of *origin-destination pairs* (OD-pairs or traffic relation). Furthermore, we are given a finite set $\mathcal{C}$ of *travel choices*. A travel choice can be a certain ticket type, e.g., single ticket or monthly ticket. Moreover, it can be a combination of a ticket type and a number of trips, which are performed within a certain time horizon, e.g., 40 trips with a monthly ticket in one month.

Let $p_{st}^i : \mathbb{R}^n \rightarrow \mathbb{R}_+$ be the *price function* for travel choice $i \in \mathcal{C}$ and OD-pair $(s, t) \in D$, i.e., it determines the price for the given travel choice and the given OD-pair. The price function depends on a *fare vector* $\mathbf{x} \in \mathbb{R}_+^n$ of $n \in \mathbb{N}$ fare variables $x_1, \dots, x_n$, which we call *fares* in the following. The *demand functions* $d_{st}^i(\mathbf{x})$ measure the amount of passengers that travel from s to t with travel choice i , depending on the fare system $\mathbf{x}$; they are assumed to be *nonincreasing*.

We denote by $\mathbf{d}_{st}(\mathbf{x})$ the vector of all demand functions associated with OD-pair (s, t) and by $\mathbf{d}(\mathbf{x}) = (\mathbf{d}_{st}(\mathbf{x}))$ the vector of all demand functions. Analogous notation is used for $(p_{st}^i(\mathbf{x}))$. The *revenue* $r(\mathbf{x})$ can then be expressed as

$$r(\mathbf{x}) := \mathbf{p}(\mathbf{x})^T \mathbf{d}(\mathbf{x}) = \sum_{i \in \mathcal{C}} \sum_{(s, t) \in D} p_{st}^i(\mathbf{x}) \cdot d_{st}^i(\mathbf{x}).$$

With this notation our general model for the fare planning problem is:

$$\begin{aligned} \text{(FPP)} \quad & \max \mathbf{p}(\mathbf{x})^T \mathbf{d}(\mathbf{x}) \\ & \text{s.t. } \mathbf{x} \in P. \end{aligned} \tag{1}$$

All restrictions on the fare variables are included in the set $P \subseteq \mathbb{R}^n$. Here, one can also include social and political aspects like a minimum level of demand or a maximum level of fares.

In the model (FPP) we assume constant costs and a constant level of service. In further investigations we included costs and maximized the profit, i.e., revenue minus costs. Other possible objectives were considered as well, e.g., maximization of the demand with respect to cost recovery. The goal is to make a first step with (FPP) towards a decision support tool for optimizing fare systems. We show the practicability of (FPP) on a prototype example in Section 3.1.

3 Fare Planing with a Discrete Choice Model

Our model expresses passenger behavior in response to fares by the demand function d_{st}^i . In this section, we use discrete choice models, especially the logit model, to obtain a realistic demand function. Therefore we assume that the passengers have full knowledge of the situation and act rationally with respect to the change of the fares. A thorough exposition of discrete choice analysis and logit models can be found in Ben-Akiva and Lerman [1].

In a discrete choice model for public transport, each passenger chooses among a finite set A of *alternatives* for the travel mode, e.g., single ticket, monthly ticket, bike, car travel, etc.

We consider a time horizon T and assume that a passenger which travels from s to t performs a random number of trips $X_{st} \in \mathbb{Z}_+$ during T , i.e., X_{st} is a discrete

random variable. We assume that passengers do not mix alternatives, i.e., the same travel alternative is chosen for all trips. Furthermore, we assume an upper bound N on X_{st} . The travel choices are then $\mathcal{C} = A \times \{1, \dots, N\}$.

Associated with each travel choice $(a, k) \in \mathcal{C}$ and each OD-pair $(s, t) \in D$ is a *utility* $U_{st}^{a,k}$, which may depend on the passenger. Each utility is the sum of an observable part, the deterministic utility $V_{st}^{a,k}$, and a random utility, or disturbance term ν_{st}^a . For $(a, k) \in \mathcal{C}$ we consider the utility $U_{st}^{a,k}(\mathbf{x}) = V_{st}^{a,k}(\mathbf{x}) + \nu_{st}^a$ that depends on the fare system $\mathbf{x}$.

Assuming that each passenger chooses the alternative with the highest utility, the probability of choosing alternative $a \in A$ in case of k trips is

$$P_{st}^{a,k}(\mathbf{x}) := \mathbb{P}[V_{st}^{a,k}(\mathbf{x}) + \nu_{st}^a = \max_{b \in A} (V_{st}^{b,k}(\mathbf{x}) + \nu_{st}^b)]. \quad (2)$$

In case of the *logit model*, which introduces the Gumbel distribution for the disturbances ν_{st}^a , this probability can explicitly be computed by the formula

$$P_{st}^{a,k}(\mathbf{x}) = \frac{e^{\mu V_{st}^{a,k}(\mathbf{x})}}{\sum_{b \in A} e^{\mu V_{st}^{b,k}(\mathbf{x})}} = \frac{1}{1 + \sum_{b \in A \setminus \{a\}} e^{\mu(V_{st}^{b,k}(\mathbf{x}) - V_{st}^{a,k}(\mathbf{x}))}}. \quad (3)$$

Here $\mu > 0$ is a scale parameter for the disturbances ν_{st}^a .

We write $d_{st}^{a,k}(\mathbf{x})$ for the amount of passengers choosing $(a, k) \in \mathcal{C}$, i.e., traveling k times during T with alternative a from s to t and similarly $p_{st}^{a,k}(\mathbf{x})$ for the price of this travel. It follows that

$$d_{st}^{a,k}(\mathbf{x}) = d_{st} \cdot P_{st}^{a,k}(\mathbf{x}) \cdot \mathbb{P}[X_{st} = k] = d_{st} \cdot \frac{e^{\mu V_{st}^{a,k}(\mathbf{x})}}{\sum_{b \in A} e^{\mu V_{st}^{b,k}(\mathbf{x})}} \cdot \mathbb{P}[X_{st} = k],$$

where d_{st} is the entry of the OD-matrix corresponding to $(s, t) \in D$. The revenue can then be written as:

$$r(\mathbf{x}) = \sum_{a \in A'} \sum_{k=1}^N \sum_{(s,t) \in D} p_{st}^{a,k}(\mathbf{x}) \cdot d_{st}^{a,k}(\mathbf{x}) = \mathbf{p}(\mathbf{x})^T \mathbf{d}(\mathbf{x}),$$

where A' is the set of public transport alternatives. This formula expresses the *expected* revenue over the probability spaces of X_{st} and disturbances ν_{st}^a .

Note that $r(\mathbf{x})$ is continuous and even differentiable if the deterministic utilities $V_{st}^{a,k}(\mathbf{x})$ (and the price functions $p_{st}^{a,k}(\mathbf{x})$) have this property.

3.1 Example with Three Alternatives

In this section, we illustrate our discrete choice approach with a real world example for the city of Potsdam. We want to optimize the current fare system including a single ticket (S) and a monthly ticket (M) for two different tariff-zones. The third travel alternative is the car (C), i.e., $A = \{M, S, C\}$. We consider a time horizon T of one month. The prices for public transport involve two fares for each tariff-zone; x_s , the single ticket fare, and x_m , the monthly ticket fare. We write $\mathbf{x} = (x_s, x_m)$ and set the prices for alternatives single, monthly ticket, and car to

$$p_{st}^{S,k}(\mathbf{x}) = x_s \cdot k, \quad p_{st}^{M,k}(\mathbf{x}) = x_m, \quad \text{and} \quad p_{st}^{C,k}(\mathbf{x}) = Q + q \cdot \ell_{st} \cdot k.$$

For alternative “car”, the price is the sum of a fixed cost Q and distance dependent operating costs q . The parameter ℓ_{st} denotes the shortest distance between s and t in kilometers for a car. We set $Q = 100$ € and $q = 0.1$ €.

We assume that the utilities are affine functions of prices and travel times t_{st}^a for traveling from s to t with alternative a . The utilities depend on the number of trips k . More precisely, we set:

$$\begin{aligned} U_{st}^{M,k}(x_M, x_S) &= -\delta_1 \cdot x_M - \delta_2 \cdot t_{st}^M \cdot k + \nu_{st}^M && \text{“monthly ticket”} \\ U_{st}^{S,k}(x_M, x_S) &= -\delta_1 (x_S \cdot k) - \delta_2 \cdot t_{st}^S \cdot k + \nu_{st}^S && \text{“single ticket”} \\ U_{st}^{C,k}(x_M, x_S) &= -\delta_1 (Q + q \cdot \ell_{st} \cdot k) - \delta_2 \cdot t_{st}^C \cdot k - y + \nu_{st}^C && \text{“car”}. \end{aligned}$$

Here, δ_1 and δ_2 are weight parameters; we use $\delta_1 = 1$ and $\delta_2 = 0.1$, i.e., 10 minutes of travel time are worth 1 €. In first computations we noticed that the behavior of the motorists could not be explained only with travel time and costs. Therefore we introduced an extra positive utility y for the car indicating the convenience of the car. We set $y \approx 93$ € for the first tariff-zone and $y \approx 73$ € for the second tariff-zone to match the current demand for the given prices in our model. The (discrete) probabilities for the number of trips are centered around 30 in an interval from 1 to $N := 60$ for all OD-pairs.

Altogether, the fare planning problem we want to consider has the form:

$$\begin{aligned} \max \quad & \sum_{k=1}^N \sum_{(s,t) \in D} d_{st} \cdot \frac{x_M \cdot e^{\mu V_{st}^{M,k}(\mathbf{x})} + x_S \cdot k \cdot e^{\mu V_{st}^{S,k}(\mathbf{x})}}{\sum_{b \in A} e^{\mu V_{st}^{b,k}(\mathbf{x})}} \cdot \mathbb{P}[X_{st} = k] \\ \text{s.t.} \quad & \mathbf{x} \geq 0 \end{aligned}$$

We set $\mu = \frac{1}{30}$. Note that the revenue function is differentiable.

The revenue function is shown on the left of Figure 1. The optimal fares for the two tariff zones are $x_s = 1.57$ (currently 1.45), $x_m = 43.72$ (32.50) for tariff-zone 1 and $x_s = 1.79$ (2.20) and $x_m = 48.21$ (49.50) for tariff-zone 2. The revenue increased by about 3% up to 2 129 971€.

3.2 Some Theoretical Results

In this section, we analyze the revenue function in case of a discrete choice demand function with a small random utility.

The second part of equation (3) emphasizes the importance of the difference of the deterministic utilities which is weighted by the parameter μ . The higher μ , the more important is the difference of the deterministic utilities for the decision, i.e., the influence of the random utility decreases. The right of Figure 1 shows a demand function for different values of μ . The choice is getting deterministic if μ tends to infinity, i.e., $P_{st}^{a,k}(x) = 1$ if a is the alternative with the maximum deterministic utility and $P_{st}^{a,k}(x) = 0$ otherwise. In this case the demand function becomes a step function.

For further analysis we omit the number of trips k and consider two travel alternatives with the following utility function $V_i^1(x) = -x$, $V_i^2(x) = -i$ for OD-pair i . The demand function for alternative 1 is

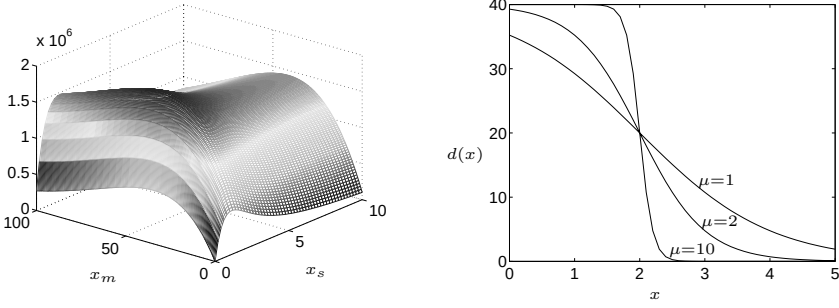


Fig. 1. *Left:* Revenue function for example single monthly ticket. *Right:* Discrete choice demand functions for different μ .

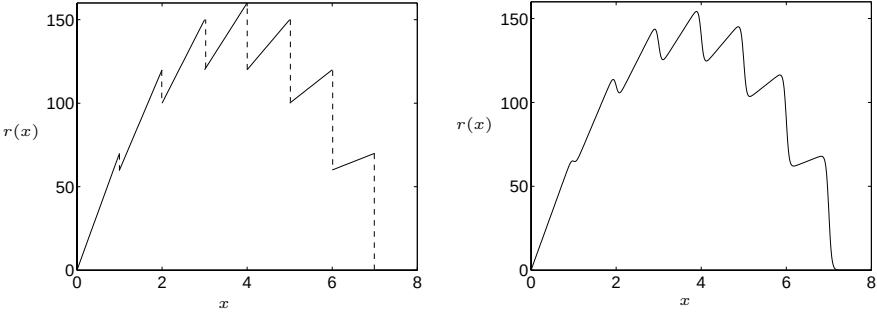


Fig. 2. *Left:* Revenue function for a discrete choice demand function for $m = 7$ OD-pairs without random utility. *Right:* Revenue function for the same example with “small” random utility ($\mu = 30$).

$$d_i(x) = d_i \cdot \frac{e^{-\mu \cdot x}}{e^{-\mu \cdot x} + e^{-\mu \cdot i}}.$$

If we set the price functions of alternative 1 to $p_i^1(x) = x$, we obtain for the revenue function for alternative 1

$$r(x) = \sum_{i=1}^m d_i \cdot \frac{e^{-\mu \cdot x}}{e^{-\mu \cdot x} + e^{-\mu \cdot i}} \cdot x.$$

$$\text{For } \mu \rightarrow \infty \quad \tilde{r}(x) := \lim_{\mu \rightarrow \infty} r(x) = \sum_{i=1}^m d_i \cdot \begin{cases} x & \text{if } x \leq i \\ 0 & \text{otherwise} \end{cases}.$$

For $x = i, i = \{1, \dots, m\}$ the revenue function $\tilde{r}$ is not continuous and has m local maxima, see left of Figure 2. This means, that if the deterministic utility approximates the utility of the alternative quite well (the random utility is small), the revenue function has m local maxima, see right of Figure 2. It is likely to construct examples with upto m^n local maxima in case of n fare variables. Therefore, the better the utilities are known, the closer the demand function is to reality. On the other hand, more local optima can appear and the problem may be hard to solve.

References

1. Moshe Ben-Akiva and Steven R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, 2. ed., MIT Press, Cambridge, 1985.
2. Ralf Borndörfer, Marika Neumann, and Marc E. Pfetsch, *Fare planning for public transport*, Report 05-20, ZIB, 2005, <http://www.zib.de/Publications/abstracts/ZR-05-20>.
3. Ralf Borndörfer, Marika Neumann, and Marc E. Pfetsch, *Optimal fares for public transport*, Proc. of Operations Research 2005, Springer-Verlag, 2005, Erscheint 2006.
4. John F. Curtin, *Effect of Fares on Transit Riding*, Highway Research Record 213 (1968).
5. Bruno De Borger, Inge Mayeres, Stef Proost, and Sandra Wouters, *Optimal Pricing of Urban Passenger Transport, A Simulation Exercise for Belgium*, Journal of Transport Economics and Policy **30** (1996), no. 1, 31–54.
6. H.W. Hamacher and A. Schöbel, *Design of zone tariff systems in public transportation*, Oper. Res. **52** (2004), no. 6, 897–908.
7. George Kocur and Chris Hendrickson, *Design of Local Bus Service with Demand Equilibration*, Transportation Science **16** (1982), no. 2, 149–170.
8. Pral Andreas Pedersen, *On the Optimal Fare Policies in Urban Transportation*, Transportation Research Part B **33** (2003), no. 37, 423–435.

Plenary and Semi-Plenary Talks

Recent Advances in Robust Optimization

Aharon Ben-Tal

MINERVA Optimization Center, Technion, Haifa, Israel
`abental@ie.technion.ac.il`

We will briefly survey the state of the art of the Robust Optimization (RO) methodology for solving convex conic optimization problems, both static and dynamic (multi-stage) emphasizing issues of computational tractability, and probabilistic guarantees satisfied by the optimal robust solution. We then introduce a recent extension of the methodology in which the solution is required to exhibit a controlled deterioration in the performance for uncertain data outside the nominal uncertainty set. Finally we discuss uncertainly affected linear control systems and introduce a novel reparameterization scheme that converts the, otherwise nonconvex, control problem into a convex programming one.

Neuro-Dynamic Programming: An Overview and Recent Results

Dimitri P. Bertsekas

Massachusetts Institute of Technology, Cambridge, U.S.A.
dimitrib@mit.edu

Neuro-dynamic programming is a methodology for sequential decision making under uncertainty, which is based on dynamic programming. The key idea is to use a scoring function to select decisions in complex dynamic systems, arising in a broad variety of applications from engineering design, operations research, resource allocation, finance, etc. This is much like what is done in computer chess, where positions are evaluated by means of a scoring function and the move that leads to the position with the best score is chosen. Neuro-dynamic programming provides a class of systematic methods for computing appropriate scoring functions using approximation schemes and simulation/evaluation of the system's performance.

Following an overview of this methodology, we will focus on recent work, which aims to approximate the cost function $J_\mu(i)$ of a policy μ with a parametric architecture of the form $\tilde{J}(i, r)$, where r is a parameter vector. This approximation may be carried out repeatedly, for a sequence of policies, in the context of a policy iteration scheme. We discuss two types of methods:

(a) **Direct methods**, where we use simulation to collect samples of costs for various initial states, and fit the architecture $\tilde{J}$ to the samples through some least squares problem. This problem may be solved by several possible algorithms, but we will focus on gradient methods.

(b) **Indirect methods**, where we obtain r by solving an approximate version of Bellman's equation

$$\tilde{J} = \tilde{T}\tilde{J},$$

where $\tilde{T}$ is an approximation to the DP mapping T_μ , designed so that the above equation has a solution. We will focus on linear architectures, where $\tilde{J}$ is of the form Φr , and Φ is a matrix whose columns can be viewed as basis functions. In this case, we obtain the parameter vector r by solving the equation

$$\Phi r = \Pi T(\Phi r), \tag{1}$$

where Π denotes projection with respect to a suitable norm on the subspace of vectors of the form Φr , and T is either the mapping T_μ or a related mapping, which also has J_μ as its unique fixed point. We can view Eq. (1) as a form of *projected Bellman equation*. We will discuss two methods, which use simulation and can be shown to converge under reasonable assumptions, and to the same limit [the unique

solution of Eq. (1)]. They both depend on a parameter $\lambda \in [0,1]$, which enters in the definition of the mapping $\tilde{T}$. A third method, which also converges to the same limit, but is much slower, is the classical TD(λ) method.

(1) *LSPE*(λ) or *least squares policy evaluation method*. This algorithm is based on the idea of executing value iteration within the lower dimensional space spanned by the basis functions. It has the form

$$\Phi r_{k+1} = \Pi T(\Phi r_k) + \text{simulation noise},$$

i.e., the current value iterate $T(\Phi r_k)$ is projected on S and is suitably approximated by simulation. The simulation noise tends to 0 asymptotically, so the method converges to the solution of the projected Bellman equation (1).

(2) *LSTD*(λ) or *least squares temporal differences method*. This algorithm computes and solves a progressively more refined simulation-based approximation to the projected Bellman equation (1). Its rate of convergence is comparable to that of LSPE(λ).

This research is contained in a series of papers by the author and his collaborators:

References

1. Bertsekas, D. P., and Ioffe, S. (1996) "Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming," Lab. for Info. and Decision Systems Report LIDS-P-2349, Massachusetts Institute of Technology.
2. Nedić, A. and Bertsekas, D. P. (2003) "Least-Squares Policy Evaluation Algorithms with Linear Function Approximation," J. of Discrete Event Systems, Vol. 13, pp. 79-110.
3. Bertsekas, D. P., Borkar, V., and Nedić, A. (2004) "Improved Temporal Difference Methods with Linear Function Approximation," in Learning and Approximate Dynamic Programming, by J. Si, A. Barto, W. Powell, (Eds.), IEEE Press, N. Y.
4. Yu, H., and Bertsekas, D. P. (2006) "Convergence Results for Some Temporal Difference Methods Based on Least Squares," Lab. for Information and Decision Systems Report 2697, MIT.

An extensive development can be found in the forthcoming book:

Dimitri Bertsekas, *Dynamic Programming and Optimal Control*, Vol. II, 3rd Edition, Athena Scientific, Belmont, MA, Dec. 2006.

Basel II – Achievements and Challenges

Klaus Duellmann

Deutsche Bundesbank, Frankfurt, Germany*

`klaus.duellmann@bundesbank.de`

1 The Goals: Risk Sensitivity and Qualitative Supervision

In late June 2004, the Basel Committee on Banking Supervision approved and published a document entitled “International Convergence of Capital Measurement and Capital Standards: A revised Framework”, better known as the “Basel II framework”. The publication of this document marked the final milestone of a process that was over five years in the making. The fundamental improvements over the Basel I Accord of 1988 help explain this relatively long time span. The main objectives of the new framework, stated by the Basel Committee in its June 1999 Consultative Paper, were the following:

- To promote safety and soundness in the financial system.
- To enhance competitive equality.
- To adopt a more comprehensive approach to addressing risks.
- To continue to focus on internationally active banks, although the new framework’s principles should also be applicable to banks of varying levels of complexity and sophistication.

Whereas the first two goals pick up where the Basel I Accord left off, the last two represent important advancements. The desire to develop a more comprehensive approach was a direct consequence of recognizing that the current regime lacks risk sensitivity in its minimum capital requirements and encourages market participants to exploit mechanisms of regulatory capital arbitrage. Capital arbitrage occurs if regulatory capital requirements do not reflect the true risk of a certain business transaction and set an incentive to pursue this transaction instead of another which is less risky but more costly in terms of regulatory capital. Mounting concerns about insufficient risk sensitivity were spurred by the creation of new instruments, an example being the triumphant success of the market for securitizations. Furthermore, advances in risk management, in particular for credit risk, which occurred in the 1990s drove a widening wedge between regulatory capital, still based on the rigid risk weight scheme of Basel I, and the bank’s own internal economic capital.

* The views expressed herein are my own and do not necessarily reflect those of the Deutsche Bundesbank.

The new framework, however, goes beyond improving the risk sensitivity of the regulatory capital calculation. It was planned from the beginning to rest on three pillars of which the minimum capital requirements are only the first.

The second pillar comprises the supervisory review process and has at its core four principles:

1. Banks should have an adequate internal capital assessment process.
2. This process should be reviewed by supervisors, who should take appropriate action if necessary.
3. Banks should be expected to operate above the minimum capital ratios.
4. Supervisors should intervene at an early stage to prevent capital from falling below the minimum levels.

The third pillar aims at strengthening market discipline through requiring the disclosure of important information about banks' capital and their riskiness.

In spite of the fundamental advancements in pillar 1, it has been argued that pillar 2 actually constitutes the most important innovation of Basel II. The reason is that it marks the turning point in the process of moving away from calculating regulatory capital as a mechanical exercise toward a qualitatively oriented supervisory regime that does more justice to the complexity and sophistication of a bank's operations.

Pillar 1 offers banks a choice between three different approaches. The revised standardized approach inherits many elements of the static risk weight scheme of Basel I, but also brings important innovations, for example, in the recognition of credit risk mitigation techniques and a previously not existing capital charge for operational risk. The Internal Ratings Based (IRB) approach is more risk sensitive and comes in two flavors, the Foundation IRB (FIRB) and the Advanced IRB (AIRB) approach. Both approaches mainly differ in the set of risk components which need to be estimated by the bank instead of being defined in the Basel II framework. In the following we focus only on the two IRB approaches.

The differentiation between the three approaches in pillar 1 reflects the idea of an evolutionary framework. Banks should select their approach in accordance with the complexity of their business operations and the sophistication of their internal risk assessment processes and, when appropriate, move to a more risk sensitive approach. Therefore, Basel II is less prescriptive and much more flexible than the previous regime.

In the following, the focus is on the first pillar, in which the capital requirements for banks' credit portfolios serve as an example of the advancements due to the Basel II framework as well as the challenges that remain.

2 The Achievements: Model-Based Risk Weight Functions

Model-based risk weight functions for a bank's credit portfolio which are a key characteristic of both IRB approaches are a fundamental change from Basel I. In the current regulatory regime, differences in credit risk are captured by risk weights dependent on a broad categorization of borrowers, e.g. OECD banks or residential mortgages. More specifically, the risk weight functions for credit risk in the new

framework are based on a single-factor default-mode asset value model which can be regarded as a highly stylized version of the more general best-practice models currently being applied in the industry. In the following we refer to this model as the Asymptotic Single Risk Factor (ASRF) model for reasons which will soon become clear. In this model the standardized asset value return X_i of firm i depends on a single systematic risk factor Y and an idiosyncratic risk factor ϵ_i :

$$X_i = \sqrt{\rho}Y + \sqrt{1 - \rho}\epsilon_i. \quad (1)$$

The parameter ρ denotes the asset correlation between any pair of firms. Since Y and ϵ_i are pairwise independent and standard normally distributed, so is X_i . Default occurs if X_i falls below a default threshold γ_i at the end of a risk horizon of one year. The (unconditional) probability of default is given by $\Phi(\gamma_i)$, where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function.

Banks hold capital as a cushion against Unexpected Loss (UL) since this source of risk can only partly be eliminated by better diversifying among borrowers. UL is defined as the difference between an adverse quantile of the portfolio's loss distribution (in Basel II at a confidence level of 99.9%) and the expected loss, namely the first moment of the loss distribution. Expected losses are traditionally covered by reserves and also considered in the calculation of the interest rate margin.

Credit risk models usually lack portfolio invariance, a convenient property for regulatory capital purposes. This property means that the UL contribution for an individual asset to portfolio UL depends only on its own risk characteristics and not on other credit-risky assets or their distribution in the portfolio.

In order to ensure portfolio invariance, the ASRF model requires two assumptions, namely that all systematic risk is driven by a single systematic factor and that the portfolio is infinitely granular.¹ If these two assumptions hold, UL is asymptotically given by the following formula:²

$$UL^{portfolio} = \sum_{n=1}^N EAD_n LGD_n \left[\Phi \left(\frac{\Phi^{-1}(PD_n) + \sqrt{\rho}\Phi^{-1}(q)}{\sqrt{1 - \rho^2}} \right) - PD_n \right] \quad (2)$$

where N denotes the number of assets in the portfolio. The ASRF model's property of portfolio invariance ensures that the contribution of each exposure to the minimum regulatory capital can be computed only from three exposure-dependent risk components, without considering other assets in the portfolio. These risk components are the Exposure At Default (EAD), the Probability of Default (PD), and the Loss Given Default (LGD) which generally need to be estimated by the bank.³

¹ See [4].

² Formula 2 differs from the IRB risk weight functions in that it does not have the scaling factor of 12.5 and the maturity adjustments which apply for maturities beyond one year. Furthermore, in the IRB model the asset correlation ρ depends on the borrower PD, and for small and medium-sized enterprises in the corporate asset class it depends also on firm size.

³ In the FIRB approach, LGDs for non-retail asset classes are defined in the Basel II framework, dependent on seniority and type of collateral, as are conversion factors for EAD of undrawn lines. For presentational purposes I do not consider maturity adjustments here, which require maturity as the fourth risk component in the AIRB approach.

In real bank portfolios, neither assumption is generally fully met. Real portfolios are not infinitely granular, nor is systematic risk from an unbalanced portfolio distribution across geographical regions and/or business fully eliminated.

In summary, an important achievement of the IRB approach is to apply for the first time model-based risk-weight functions which measure credit risk much finer than the relatively coarse risk-weight scheme of Basel I. The new model-based approach, however, also creates two challenges, namely to account for violations of the granularity assumption and of the single-risk factor assumption if necessary.

Identifying a violation of these two assumptions, namely single-name and sectoral concentration as sources of credit concentration risk, suggests that this risk is not a different risk category but rather an important driver of credit risk in general. Treating both risk types as separate forms of credit risk becomes useful in the context of the ASRF model since this model by construction does not take them into account.

3 The Challenge: Measuring Concentration Risk

Before discussing how single-name and sector concentration risk can be measured, it is useful to briefly consider their empirical relevance. [8] from S&P carry out an analysis for the 100 largest, externally rated Western European banks. The ratio of a bank's twenty largest exposures relative to its capital base is used as a yardstick to measure concentration risk. The authors find great diversity among banks and "sometimes exceedingly-high [single-name] concentrations".

[2] consider the impact of sector concentration and single-name concentration on portfolios of real banks. They use a standard CreditMetrics default-mode portfolio model⁴ and differentiate between 11 business sectors. The sector definitions are based on the Global Industry Classification Standard (GICS), launched by Standard&Poor's and Morgan Stanley Capital International in 1999. For the measurement of sector concentration they construct a sequence of portfolios in which the relative share of exposures in a particular sector is subsequently increased. This sequence starts with a portfolio reflecting the aggregate sector distribution of the German banking system and ends with a portfolio that is concentrated in a single sector. Information about the sectoral distribution of exposures is based on a sample of 2224 banks and was extracted from the central credit register maintained at the Deutsche Bundesbank. Inter-sector correlations were estimated from MSCI stock index returns.

The authors find that in the case of a single-sector portfolio the increase in UL from the benchmark portfolio is 50%. For the next sector-wise most concentrated of their portfolios the corresponding increase is still 40% and this portfolio closely resembles a real bank's portfolio in terms of its exposure distribution across sectors. They also measure the impact of granularity as 7.5%, yet this is a conservative estimate given that it is based on subsets of the banks' portfolios which comprise only exposures of at least 1.5m euros. Therefore, the diversification effect from smaller, particularly retail exposures is not accounted for. These results suggest that the intuition of sectoral concentration generally being more important than single-name concentration applies.

⁴ See [6].

The rest of this section briefly sketches three methodologies to measure credit concentration risk. The focus is on practical approaches rather than on giving an exhaustive overview of existing models.

From a methodological perspective the infinite-granularity assumption poses the smaller challenge of the two types of concentration risk. An extension of the single-factor model which accounts for granularity was already proposed in the Second Consultative Paper⁵ in 2001 and later refined by [5], in particular in terms of its applicability. This granularity adjustment is given in closed form and mainly requires the inputs which are already needed to compute UL. The biggest challenge for its implementation is arguably the grouping of single-exposures into larger risk entities. This is important for risk measurement purposes because default events are borrower-specific and often borrowers are so closely linked that a default would spread out to the others. The mathematical structure of the granularity adjustment is simple and in the special case of a homogenous portfolio in terms of PD, LGD, and maturity, it collapses to a linear function of the Herfindahl-Hirschman Index, calculated as the sum of squared relative exposures in the portfolio.

The one-factor assumption, however, poses a significant challenge since it constitutes the core of the ASRF model. Multi-factor models appear appropriate for measuring sector concentration but impose a computational burden since UL usually has to be computed by Monte Carlo simulation. Furthermore, they account for risk concentration but not as a separately identifiable part of the marginal risk contribution of single exposures, so they do not provide an explicit capital figure for the contribution of single-name concentration or sectoral concentration in the portfolio. For practical purposes, it would be useful to have an analytic UL approximation which combines the risk-sensitivity of a multi-factor structure with a high degree of tractability. To my knowledge there are currently two proposals which look promising.

The first approach by [3] is based on the idea of adjusting UL from the ASRF model by a multiplicative *diversification factor* which depends on a measure of dispersion of exposures across sectors and a central tendency measure of sector correlations. The authors estimate a parametric surface of the diversification factor by extensive Monte Carlo simulations. The parameters are set to minimize the error if UL is measured by the ASRF model, extended by the diversification factor, instead of by the multi-factor model.

The second proposal by [2] uses an analytic VaR approximation, presented first by [7]. Instead of applying this approximation formula on the borrower level, as proposed in [7], they aggregate the inputs on a sector level and use an exposure-weighted PD and the aggregate exposure of each sector instead. In this way they further reduce the computational burden. They find that the approximation performance is extremely good for portfolios with homogeneous PDs, assuming that idiosyncratic risk is fully diversified.

The methodologies of Garcia Cespedes et al. and Duellmann and Masschelein both focus on sectoral concentration and do not explicitly account for heterogeneity on the borrower level. Duellmann and Masschelein carry out various robustness checks of their results with portfolios characterized by heterogeneous PDs and unbalanced single-exposure distributions inside sectors. They find that the errors from using an average PD and from ignoring exposure heterogeneity partly balance each

⁵ See [1].

other. Using average PDs *ceteris paribus* overestimates the “true” VaR, whereas ignoring granularity *ceteris paribus* underestimates it.

Since the first of the two effects is stronger for representative loan portfolios, the authors conclude that their approximation formula is conservative in such cases. Furthermore, the approximation quality is satisfactory, given that the relative error is less than 10% in absolute terms. Pending further robustness checks, their findings show a promising avenue of research into how UL can be estimated by analytic approximations reasonably well and with quite parsimonious data requirements.

4 Summary and Conclusions

The new framework of Basel II marks an important milestone in the development of banking supervision for various reasons. Some of its elements, such as the three-pillar structure and the turning to a qualitatively oriented supervisory regime, have been touched on in this article, while others, such as the new capital charge for operational risk, had to be excluded. One of its most important achievements is the increase of risk sensitivity of regulatory minimum capital requirements in pillar I. This achievement was discussed in the context of banks’ loan portfolios which are arguably the most important part of traditional banking business.

Portfolio invariance, although desirable for the purpose of regulatory capital requirements, motivates two assumptions of the model, which can have great practical relevance. These are the assumptions of infinite granularity and that all systematic risk is driven by a single systematic risk factor, loosely speaking that the portfolio has neither single-name concentration nor sectoral concentration. In recognition of their practical relevance, credit concentration risk has been included in pillar 2 as a risk which is not fully captured by pillar 1.

Various approaches to measuring credit concentration risk have been discussed in this paper. Notwithstanding promising first results in assessing single-name and sector concentration, further work is warranted to obtain improved methods which are both tractable and more risk sensitive by accounting for single-name and in particular for sectoral concentration risk.

References

1. Basel Committee on Banking Supervision. The New Basel Capital Accord, Consultative Document. <http://www.bis.org/publ/bcbsca03.pdf>, 2001.
2. K. Duellmann and N. Masschelein. Sector concentration in loan portfolios and economic capital. Deutsche Bundesbank Discussion Paper, Series 2, No. 9, and NBB working paper, 2006.
3. J. C. Garcia Cespedes, J. A. de Juan Herrero, A. Kreinin, and D. Rosen. A simple multi-factor “factor adjustment” for the treatment of diversification in credit capital rules. *Journal Credit Risk*, 2(3):57–85, 2006.
4. M. Gordy. A comparative anatomy of credit risk models. *Journal of Banking and Finance*, 24:119–149, 2000.
5. M. Gordy and E. Luetkebohmert. Granularity adjustment for Basel II. Unpublished Working Paper, 2006.

6. G.M. Gupton, C.C. Finger, and M. Bhatia. *CreditMetrics – Technical Document*. New York: Morgan Guaranty Trust Co., 1997.
7. M. Pykhtin. Multi-factor adjustment. *Risk Magazine*, 17(3):85–90, 2004.
8. P. Tornqvist and B. de Longevialle. Research: Concentration risks remain high at european banks. Standard&Poor’s Ratings Direct 22-Oct-2004, 2004.

How to Model Operational Risk If You Must

Paul Embrechts

ETH Zurich, Switzerland
`embrechts@math.ethz.ch`

Both under Solvency 2 and Basel II, operational risk is an important risk category for which the financial industry has to come up with a capital charge. Under Basel II, Operational Risk is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk. In this talk I will discuss some of the issues underlying the quantitative modelling of operational risk.

References

1. McNeil, A., Frey, R. and Embrechts, P. *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press, 2005.
2. Neslehova, J., Embrechts, P. and Chavez-Demoulin, V. Some issues underlying the AMA modeling of operational risk, ETH Zurich, preprint, 2006.

Integer Quadratic Programming Models in Computational Biology

Harvey J. Greenberg

Department of Mathematical Sciences, University of Colorado at Denver and Health Sciences Center, U.S.A.
`harvey.greenberg@cudenver.edu`

1 Introduction

This presentation has two purposes: (1) show operations researchers how they can apply quadratic binary programming to current problems in molecular biology, and (2) show formulations of some combinatorial optimization problems as integer programs. The former purpose is primary, and I wish to persuade researchers to enter this exciting frontier. The latter purpose is part of a work in progress.

I begin with some background, particularly in the biology; mathematical programming terms can be found in the *Mathematical Programming Glossary* [10]. I then present four integer quadratic programming (IQP) models. My last section indicates some avenues for research.

2 Background

In this section I introduce a few things about the underlying biology, which are needed to get started. I then give the general form of the underlying model and discuss some of its attributes. In both cases, I elaborate in the next section in the context of each model.

2.1 Molecular Biology

All life depends upon three critical molecules:

1. DNA, which contains information about how a cell works
2. RNA, which makes proteins and performs other functions
3. Proteins, which are regarded as the workers of the cell.

DNA is a double-stranded sequence of nucleic acids: Adenine, Cytosine, Guanine, and Thymine. RNA is a single-stranded sequence of nucleic acids: Adenine, Cytosine, Guanine, and Uracil. The genetic code maps each triple of nucleic acids into one of 20 amino acids. A *peptide bond* is a bonding of two amino acids. A protein is determined by a sequence of successively bonded amino acids. Figure 1 shows the

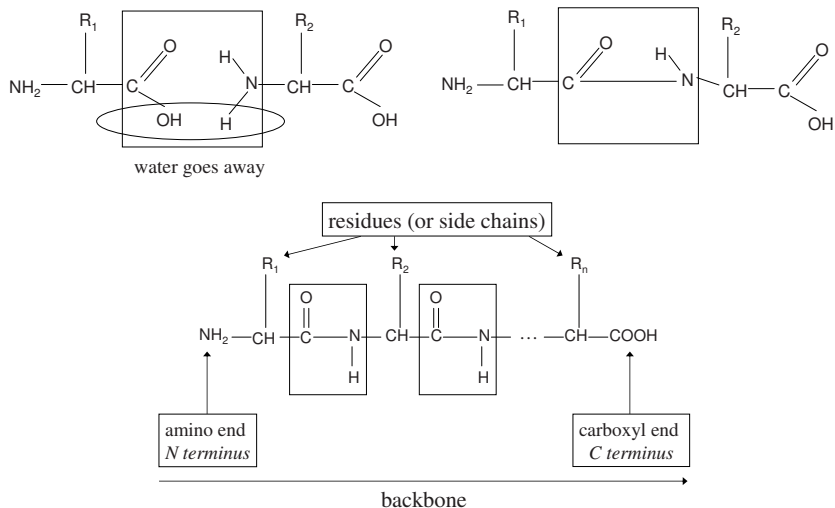


Fig. 1. Bonding Amino Acids

bonding of two amino acids — the carboxyl end of one bonds with the amino end of the other, discarding water. It then shows a generic backbone defined by a sequence of amino acids.

The *central dogma* is the information flow from DNA to proteins. (This use of “dogma” may seem strange — see [5, p. 116] for an explanation.) Figure 2 illustrates this. I shall say more as needed, in the context of each problem.

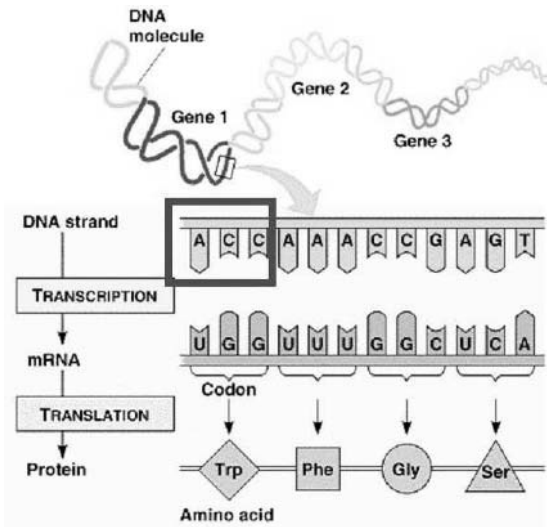


Fig. 2. The Central Dogma of Molecular Biology

2.2 Integer Quadratic Programming

The IQP model has the form:

$$\text{opt } x'Qx + dy : x \in \{0,1\}, x \in X(y), y \in Y,$$

where y are continuous variables (which may be absent), restricted to the polyhedron, Y , and $X(y)$ restricts the binary decision variable (x) to a polyhedron that depends upon y . (Think of this as some system like $Ax + By \leq b$.)

One approach is to *linearize* the objective function by introducing auxiliary variables, w , replacing the quadratic form with cw . There are different ways to do this, and I refer you to Adams, et al. [1]. Another approach is to convexify the quadratic form by adding or subtracting diagonal terms from the fact that $x^2 = x$ for $x \in \{0,1\}$. In particular, we can choose λ such that adding $\sum_j \lambda_j (x_j^2 - x_j)$ to the objective renders the quadratic form positive definite. Just as there are variations in linearization, so there are algorithm design choices in the convexification [3, 8] (e.g., choosing λ).

This presentation is about the models, not the algorithms to solve them. All of the models I present have the following properties.

- They are NP-complete.
- Some special cases have polynomial algorithms.
- Some have approximation algorithms.
- Some have been analyzed for deep cuts and pre-processing.
- Very few have been analyzed for linearization.
- None have been analyzed for convexification.

One may ask, “Why use a general IQP approach?” This must take more computer time and space to solve it, compared to a special-purpose algorithm for each particular problem. My answer is that we gain the following:

Understanding. Seeing the algebraic structure of a model offers a vantage for insights not necessarily obtained from a purely combinatorial vantage.

Extension. Often, we want to add constraints or otherwise extend the basic model. This is easy to do with an IQP approach, but it might require an entirely new algorithm design for a direct, problem-specific approach.

Management. It is easier to manage an IQP model, particularly with good management practices, such as separation of model specification from data that determines *instances* of the model.

Transfer. We can compare IQP models of different biology problems and gain insights into all of them that we might otherwise miss.

3 Models

In this section I present four IP models with quadratic objectives; other recent surveys with linear objectives can be found in [4, 7]

3.1 Multiple Sequence Alignment

Two fundamental biological sequences are taken from the alphabet of nucleic acids, $\{a, c, g, t\}$ or from the alphabet of amino acids, $\{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$. The former are segments of DNA (or RNA if t is replaced by u); the latter are segments of proteins.

The biology problem is to seek similarities among a given set of sequences from the same alphabet. This might be to:

- Understand life through evolution
- Identify families of proteins to infer structure or function from sequence
- Diagnose disease
- Retrieve similar sequences from databases.

The problem is called *Multiple Sequence Alignment* (MSA).

An early application of MSA that illustrates its importance is given by Riordan, et al. [14], who discovered the Cystic Fibrosis Transmembrane Regulator gene and its connection to Cystic Fibrosis. Even before then, algorithms for MSA [13, 15] were developed for the biologists who used computational methods for their research.

One key to defining an objective is the notion of a *gap*. This is a sequence of insertions or deletions (called *indels*) that occur during evolution. For example, if one applies a simple Hamming distance to the sequences, **acacta** and **tacact**, they are six characters apart. However, inserting gaps we obtain the *extended sequences*,

```
-acacta
  ||||
tacact-
```

This has a Hamming distance of only two. The middle sequence, **acact**, is the same in the alignment. We may suppose there was a common ancestor that inserted **a** at the end to produce the first sequence and deleted **t** at the beginning to obtain the second.

Two sequences can be optimally aligned by dynamic programming, where “optimal” is one that maximizes an objective that has two parts:

1. a *scoring function*, given in the form of an $m \times m$ matrix S , where m is the size of the alphabet. The value of S_{ij} measures a propensity for the i^{th} alphabet-character in one sequence to align with the j^{th} alphabet-character in some position of the other sequence.

Example: Let $s = \mathbf{agg}$ and $t = \mathbf{gac}$. In the direct alignment, the total score is $S_{ag} + S_{ga} + S_{gc}$.

2. a *gap penalty function*, expressed in two parts: a “fixed cost” of beginning a gap, denoted G_{open} , and a cost to “extend” the gap, denoted G_{ext} .

Example: Let $s = \mathbf{agg}$ and $t = \mathbf{cgag}$. One alignment is to put a gap in the first sequence:

```
ag-g
cgag
```

Figure 3 shows four different alignments for the two nucleic acid sequences, **agt** and **gtac**. Suppose the scoring matrix is

$\begin{matrix} & \text{a} & \text{c} & \text{g} & \text{t} \\ \text{a} & 2 & -1 & -2 & 0 \\ \text{c} & -1 & 2 & 0 & -2 \\ \text{g} & -2 & 0 & 2 & -1 \\ \text{t} & 0 & -2 & -1 & 2 \end{matrix}$

$$S = \begin{bmatrix} 2 & -1 & -2 & 0 \\ -1 & 2 & 0 & -2 \\ -2 & 0 & 2 & -1 \\ 0 & -2 & -1 & 2 \end{bmatrix}.$$

Then, the scores (without gap penalties) are 4, 0, and -3 , respectively.

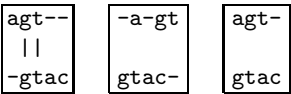


Fig. 3. Three Alignments for Two Sequences

The total objective function for the 2-sequence alignment problem has the form

$$\sum_{i,j} S_{s_i t_j} - G_{\text{open}}(N_s + N_t) - G_{\text{ext}}M_s + M_t),$$

where the sum is over aligned characters, s_i from sequence s with t_j from sequence t . The number of gaps opened in sequence s is N_s , and it is N_t in sequence t . The number of gap characters ($-$) is M_s in sequence s and M_t in sequence t . In the example of fig. 3, if $G_{\text{open}}=2$ and $G_{\text{ext}}=1$, the gap penalties are 7, 9, and 3, respectively.

Subtracting the gap penalties from the scores, the objective values of the alignments are -3 , -9 , and -6 , respectively. The first alignment is (uniquely) optimal.

There are different scoring methods, but this is the common one, which I use. One way to evaluate an MSA is by summing pairwise scores. Figure 4 shows an example. Using the same scoring matrix as above, the sum-of-pairs score is shown for each column. For example, column 1 has $3S_{aa} + 3S_{ac} = 3$. The sum of pairwise scores for column 2 is zero because we do not score the gaps by columns; they are penalized for each sequence (row of alignment). The total objective value is $31 - 28 = 3$.

	Gap penalty
a-gagt-act---	8
aagtat--at---	7
a--tataa----t	8
c-gta--actcct	5
score: 3066020606002	- 28 = Total
Total = 31	

Fig. 4. A Multiple Alignment of Four Sequences

The IQP model is as follows. Let

$$\begin{aligned}
x_{i\ell k} &= \begin{cases} 1 & \text{if } i^{\text{th}} \text{ character of sequence } \ell \text{ is assigned to column } k; \\ 0 & \text{otherwise.} \end{cases} \\
y_{\ell k} &= \begin{cases} 1 & \text{if sequence } \ell \text{ has a gap in column } k; \\ 0 & \text{otherwise.} \end{cases} \\
z_{\ell k} &= \begin{cases} 1 & \text{if sequence } \ell \text{ opens a gap in column } k; \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Then, the IQP model for the MSA of sequences $s^1, \dots, s^m$ is given by:

$$\begin{aligned}
&\max \sum_k \sum_{\ell, \ell' > \ell} \sum_{i, j} S_{s_i^{\ell'} s_j^{\ell}} x_{i\ell k} x_{j\ell' k} - \left(\sum_{k, \ell} G_{\text{open}} z_{\ell k} + G_{\text{ext}} y_{\ell k} \right) \\
&\text{subject to:} \\
&\quad \sum_k x_{i\ell k} = 1 \quad \forall i, \ell \\
&\quad \sum_{k' > k} x_{i+1, \ell k'} \geq x_{i\ell k} \quad \forall i < L_{\ell}, \ell, k \\
&\quad y_{\ell k} + \sum_i x_{i\ell k} = 1 \quad \forall k, \ell \\
&\quad y_{\ell k} - y_{k-1, \ell} - z_{\ell k} \leq 0 \quad \forall k, \ell \\
&\quad x \in \{0, 1\}, 0 \leq y, z \leq 1.
\end{aligned}$$

I do not explicitly require y, z to be binary; that is implied by x binary for basic solutions. (An interior solution yields fractional values of z if $G_{\text{open}} = 0$.)

The first sum in the objective is the sum-of-pairs score, from which we subtract the total gap penalty. The first constraint requires that each character (i) in each sequence (ℓ) is assigned to some column (k). The second constraint preserves the character order of each sequence — if the i^{th} character of string ℓ is assigned to column k ($x_{i\ell k} = 1$), then its successor ($i+1$) must be assigned to a subsequent column ($k' > k$). The third constraint requires that, for each column of each sequence, either a character is assigned or it is in a gap. Finally, the last constraint requires that a gap is opened (i.e., $z_{\ell k}$ is forced to 1) if it changed from no gap assignment ($y_{\ell k-1} = 0$) to a gap assignment ($y_{\ell k} = 1$).

3.2 Lattice Protein Folding

We are given the amino acid sequence of a protein, and we want to know its structure after it *folds* — that is, twists and turns to establish atomic bonds. Structures include the *alpha helix* and *beta sheet*, and these determine (in some way) the *function* of the protein. There have been many approaches to this over the past three decades, and the literature is vast. Here I consider one model introduced by Hart and Istrail [9].

Each amino acid has several properties, and one of particular importance is its *hydrophobicity*. It is hydrophobic, denoted H, if it does not like water; it is hydrophilic, denoted P, if it does. Some classes of proteins, namely the *globular* ones, have a *hydrophobic core*, and that helps to predict some of the structure.

A *lattice* is a grid approximation to space, and we want to assign amino acids to points on the grid such that the total number of hydrophilic neighbors is a maximum. Figure 5 shows an example. The score is 2, coming from the bonding of amino acid residues 1-6 and 2-5.

Hart and Istrail [9] laid the foundation for applying combinatorial optimization approximation algorithms. Their proof-technique for establishing their approximation algorithms has become a paradigm for this area.



Fig. 5. Example Lattice Fold

The IQP model is as follows. Let

$$x_{ip} = \begin{cases} 1 & \text{if acid } i \text{ is assigned to point } p; \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{H}$ denote the set of hydrophobic acids. Further, let $\mathcal{N}(p)$ denote the neighbors of point p (excluding p). For a square grid, one can use the Manhattan distance to define neighbors:

$$\mathcal{N}(p) = \{q : |X_p - X_q| + |Y_p - Y_q| = 1\},$$

where (X_p, Y_p) denotes the coordinates of point p .

Then, the IQP model for the Lattice Protein Folding Problem for a sequence of n amino acids is:

$$\begin{aligned} & \max \sum_p \sum_{q \in \mathcal{N}(p)} \sum_{i, j \in \mathcal{H} : j > i+1} x_{ip} x_{jq} \\ & \text{subject to:} \\ & \sum_p x_{ip} = 1 \quad \forall i \\ & \sum_i x_{ip} \leq 1 \quad \forall p \\ & \sum_{q \in \mathcal{N}(p)} x_{i+1, q} \geq x_{ip} \quad \forall p, i < n \\ & \sum_{q \in \mathcal{N}(p)} x_{i-1, q} \geq x_{ip} \quad \forall p, i > 1 \\ & x \in \{0, 1\}. \end{aligned}$$

The objective scores a 1 when hydrophobic acids i and j are assigned to neighboring points p and q ($x_{ip} = x_{jq} = 1$). The added condition $j > i + 1$ is to avoid double counting and exclude those that are already adjacent in the given sequence.

The first constraint requires that each acid be assigned to exactly one point; the second requires that at most one point is assigned to an acid. The last two constraints require backbone neighbors to remain neighbors in the fold — that is, if acid i is assigned to point p ($x_{ip} = 1$), acid $i \pm 1$ must be assigned to some neighbor ($x_{i \pm 1, q} = 1$ for some $q \in \mathcal{N}(p)$).

We must add *symmetry exclusion* constraints [2]. Any rigid motion, like translation and rotation, appears as an alternative optimum since the variables have different values, but such groups are the same fold. For example, the fold in fig. 5 can be rotated, as shown in fig. 6. I rotated it 90° clockwise about the middle acid (#5). Number the points such that $x_{11} = 1$ in the original fold, so $x'_{13} = 1$ in the rotation — that is, even though the folds are the “same,” the IQP solutions are different.

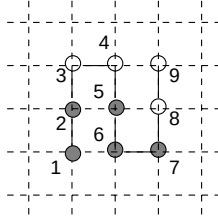


Fig. 6. Example Lattice Fold (fig. 5) Rotated 90° Clockwise

Extend the model to eliminate some symmetries at the outset with the following *symmetry exclusion constraints*:

Fix middle acid at mid-point of grid: $x_{mp_m} = 1$

Restrict acid 1 to the upper half of quadrant III: $\sum_{p \in Q} x_{1p} = 1$.

Fixing the middle acid prevents translation, and restricting acid 1 to the upper half of quadrant III, as indicated in fig. 7, prevents equivalent folds by reflection about the 45° line.

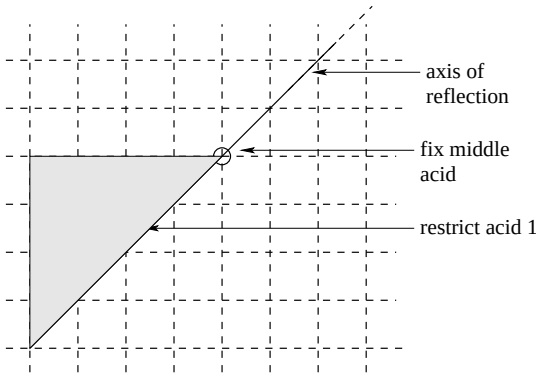


Fig. 7. Symmetry Exclusions at Outset

Other symmetries arise during branch-and-bound. Without symmetry exclusion, branch-and-bound would take unnecessary searches for what would appear to be a potentially better subtree.

The HP lattice model extends to any neighbor properties, like charge, volume, and mass. Suppose we score λ_k for neighbors such that one is in $\mathcal{I}_k$ and the other is in $\mathcal{J}_k$. The simple HP case is with $\mathcal{I}_k = \mathcal{J}_k = \mathcal{H}$ and $\lambda_k = 1$. Another interesting case is with oppositely charged amino acids, where $\mathcal{I}_k$ is the set of negatively charged acids, and $\mathcal{J}_k$ is the set of positively charged acids.

Let $\mathcal{K}$ denote the set of all such neighbor-scoring pairs. Then, the generalized model for the extended scoring alphabets is:

$$\begin{aligned}
& \max \sum_p \sum_{q \in \mathcal{N}(p)} \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{I}_k} \sum_{j \in \mathcal{J}_k: j > i+1} \lambda_k x_{ip} x_{jq} \\
& \text{subject to:} \\
& \quad \sum_p x_{ip} = 1 \quad \forall i \\
& \quad \sum_i x_{ip} \leq 1 \quad \forall p \\
& \quad \sum_{q \in \mathcal{N}(p)} x_{i+1,q} \geq x_{ip} \quad \forall p, i < n \\
& \quad \sum_{q \in \mathcal{N}(p)} x_{i-1,q} \geq x_{ip} \quad \forall p, i > 1 \\
& \quad x_{mpm} = 1, \sum_{p \in Q} x_{1p} = 1, x \in \{0,1\}.
\end{aligned}$$

3.3 Rotamer Assignments

Part of the protein folding problem is knowing the side-chain *conformations*. This means knowing the three torsion angles of the bonds. The rotation about a bond is called a *rotamer*, and there are libraries that give rotamer likelihoods for each amino acid. There are about 10 to 50 rotamers per side-chain, depending upon the amino acid and what else we know (such as whether it is located in a helix or beta sheet).

Besides its role in determining a protein's structure, the rotamer assignment problem applies to determining a min-energy docking site for a ligand (which is a small molecule, like a drug). This problem is characterized by few sites. If the protein and ligand are known, the problem dimensions are small, but if the protein or the ligand is to be engineered, there can be about 500 rotamers per site (20 acids @ 25 rotamers each). We could know a protein, making the number of rotamers per site fairly small, but the protein could have many sites (hundreds). This is the case for predicting side-chain conformations, as part of the protein structure prediction. There are other bioengineering problems associated the rotamer assignment.

Gordon, et al. [6] analyzed a standard integer linear programming model, and Kingsford, et al. [11] considered semi-definite bounding instead of LP relaxation. Xie and Sahinidis [16] applied a global optimization algorithm after reducing the problem with *Dead-End Elimination* (DEE). (That is a pre-processing technique for this particular problem that is very powerful in practice.) Here, I focus on the IQP formulation.

The IQP model is as follows. Let $r \in \mathcal{R}_i$ = set of rotamers that can be assigned to site i , and

$$x_{ir} = \begin{cases} 1 & \text{if rotamer } r \text{ is assigned to site } i; \\ 0 & \text{otherwise.} \end{cases}$$

Then, the IQP model for the Rotamer Assignment Problem is the quadratic semi-assignment problem:

$$\begin{aligned}
& \max \sum_i \sum_{r \in \mathcal{R}_i} \sum_{j > i} \sum_{t \in \mathcal{R}_j} E_{irjt} x_{ir} x_{jt} \\
& \text{subject to: } \sum_r x_{ir} = 1 \quad \forall i, x \in \{0,1\}.
\end{aligned}$$

The objective function includes two types of energy: (1) within a site ($E_{irir} x_{ir}^2 = \mathcal{E}_{ir} x_{ir}$), and (2) between rotamers of two different sites, E_{irjt} for $i \neq j$.) The summation condition $j > i$ is to avoid double counting, where $E_{irjt} = E_{jtir}$.

It is easy to add *conflict constraints* of the form $x_{ir} + x_{jt} \leq 1$, or more general logical conditions that represent the biological realities.

3.4 Protein Comparison by Contact Maps

The *contact map* of a protein is a graph, $G = [V, E]$, where V represents the amino acids that bonded to form the protein. Suppose we know its native state, and define an edge between two nodes if their distance is within some threshold, τ (typically, $3.5\text{\AA} \leq \tau \leq 5.5\text{\AA}$).

Given the contact maps of two proteins, G_1, G_2 , define their similarity to be the largest subgraphs that are isomorphic. Here, “largest” is measured by the number of edges. The Contact Map Optimization (CMO) problem is to find largest isomorphic subgraphs, constrained to have the same node ordering (i.e., preserve backbone order). Figure 8 shows an example with the optimal solution. The dashed lines indicate the seven pairs of nodes that are associated. The darkened edges are associated, for a total score of 5.

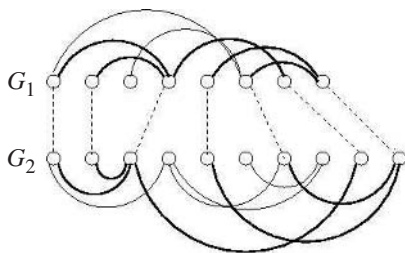


Fig. 8. Example of Two Contact Maps with Associated Nodes

A linear integer programming approach was introduced by Lancia, et al. [12]. They developed deep cuts by exploiting the problem structure.

The preservation of the backbone order is equivalent to not having any *crossing* — that is, associations $i \leftrightarrow j$ and $k \leftrightarrow \ell$ such that $i < k$ but $j > \ell$. The situation is depicted in fig. 9. Our model must disallow every crossing, but it is sufficient to eliminate all 2-crossings.

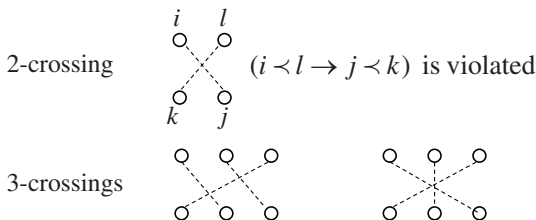


Fig. 9. Crossings to be Excluded

The IQP model is as follows. Let

$$x_{ij} = \begin{cases} 1 & \text{if node } i \in V_1 \text{ is associated with node } j \in V_2; \\ 0 & \text{otherwise.} \end{cases}$$

Then, the IQP model for the Contact Map Optimization Problem is:

$$\begin{aligned}
& \max \sum_{\substack{(i,k) \in E_1 \\ (j,\ell) \in E_2}} x_{ij} x_{k\ell} \\
& \text{subject to: } \sum_j x_{ij} \leq 1 \ \forall i, \sum_i x_{ij} \leq 1 \ \forall j \\
& x_{ij} + x_{k\ell} \leq 1 \text{ for } 1 \leq i < k < |V_1|, 1 \leq \ell < j < |V_2| \\
& x \in \{0,1\}.
\end{aligned}$$

The objective scores a 1 when edge $(i,k) \in E_1$ is associated with edge $(k,\ell) \in E_2$. That happens when the endpoint nodes are associated. The first constraint limits a node in V_1 to be associated with at most one node in V_2 ; the second is likewise. The conflict constraints preserve the backbone ordering by disallowing any 2-crossing.

The 2-crossing conflict constraints are known to be weak, compared to easily derived *stronger* systems of inequalities. Figure 10 shows an example for $|V_1| = |V_2| = 3$. The systems have the same binary solutions, and every (continuous) solution to the stronger system is a solution to the original one. However, the original system allows the fractional solution with $x_{ij} = 1/2$ for all i,j , whereas the stronger system does not admit this solution.

2-Crossing Constraints	Stronger System
$x_{12} + x_{21} \leq 1$	$x_{12} + x_{21} + x_{31} \leq 1$
$x_{12} + x_{31} \leq 1$	$x_{13} + x_{21} + x_{22} + x_{31} \leq 1$
$x_{22} + x_{31} \leq 1$	$x_{23} + x_{32} + x_{31} \leq 1$
$x_{13} + x_{21} \leq 1$	$x_{13} + x_{23} + x_{32} \leq 1$
$x_{13} + x_{31} \leq 1$	
$x_{23} + x_{31} \leq 1$	
$x_{13} + x_{22} \leq 1$	
$x_{13} + x_{32} \leq 1$	
$x_{23} + x_{32} \leq 1$	

Fig. 10. Example of Stronger No-Crossing Constraints

4 Summary and Avenues for Research

The formulations have some obvious commonalities:

- Primary binary variables are assignments.
- Constraints include [semi-]assignment equations or limits.
- There may be conflict constraints.
- There may be ordering constraints.
- Quadratic objectives represent scoring for two assignments.

These formulations need more work to capture common properties that can be integrated into the design of an algorithm using *general* techniques. This involves

Multiple Sequence Alignment $\begin{aligned} \max \quad & \sum_{s_i^{\ell'}, s_j^{\ell'}} S_{s_i^{\ell'} s_j^{\ell'}} x_{i\ell k} x_{j\ell' k} : \\ & - \sum (G_{\text{open}} z_{\ell k} + G_{\text{ext}} y_{\ell k}) \\ & \sum_k x_{i\ell k} = 1 \\ & y_{\ell k} + \sum_i x_{i\ell k} = 1 \\ & \sum_{k' > k} x_{i+1, \ell k'} \geq x_{i\ell k} \\ & y_{\ell k} - y_{k-1, \ell} \leq z_{\ell k} \end{aligned}$	MSA without gaps $\begin{aligned} \max \quad & \sum_{s_i^{\ell'}, s_j^{\ell'}} S_{s_i^{\ell'} s_j^{\ell'}} x_{i\ell k} x_{j\ell' k} : \\ & \sum_k x_{i\ell k} = 1 \\ & \sum_i x_{i\ell k} \leq 1 \\ & \sum_{k' > k} x_{i+1, \ell k'} \geq x_{i\ell k} \end{aligned}$	
Lattice Protein Folding $\begin{aligned} \max \quad & \sum x_{ip} x_{jq} : \\ & \sum_p x_{ip} = 1, \sum_p x_{ip} \leq 1 \\ & \sum_{q \in \mathcal{N}(p)} x_{i+1, q} \geq x_{ip} \\ & \sum_{q \in \mathcal{N}(p)} x_{i-1, q} \geq x_{ip} \end{aligned}$	Rotamer Assignment $\begin{aligned} \max \quad & \sum E_{irjt} x_{ir} x_{jt} \\ & \sum_r x_{ir} = 1 \\ & [x_{ir} + x_{jt} \leq 1] \end{aligned}$	Contact Map Optimization $\begin{aligned} \max \quad & \sum x_{ij} x_{kl} \\ & \sum_j x_{ij} \leq 1, \sum_i x_{ij} \leq 1 \\ & x_{ij} + x_{kl} \leq 1 \end{aligned}$

Fig. 11. All IQP Models

- Exploit structures (e.g., DEE for Rotamer Assignment)
- Exclude symmetries (e.g., rigid motion for Lattice Protein Folding)
- Linearizations — trade-off LP relaxation strength with setup computation
- Convexification — how to choose parameters
- Cut generation (e.g., stronger inequalities for Contact Map Optimization)
- Decomposition (e.g., Dantzig-Wolfe)

Finally, I want to address *uncertainty*. A part of that is to be able to identify alternative optima (or near optima), eliminating “false alternatives,” which are symmetries. The geometric symmetry in the lattice protein folding problem is easy to visualize, but many combinatorial optimization problems have symmetries, such as travelling salesman tours and graph color assignments. The biologist wants true alternative optima, such as alternative sequence alignments. In the rotamer assignment problem, the biologist would like a ranking of the minimum-energy solutions, which she could cut off above the absolute minimum. This is because the energy data has errors and some other assignment might have more significance not reflected in the model.

Acknowledgements

I am grateful for being introduced to computational biology when I visited Sandia National Laboratories in 2000. I particularly appreciate having worked with Bill Hart and Giuseppe Lancia. I also acknowledge my current collaboration with Dick Forrester, which is producing some interesting experimental results.

References

1. W.P. Adams, R.J. Forrester, and F. Glover. Comparisons and enhancement strategies for linearizing mixed 0-1 quadratic programs. *Discrete Optimization*, 1(2):99–120, 2004.
2. R. Backofen and S. Will. Excluding symmetries in constraint-based search. *Constraints*, 7(3–4):333–349, 2002.
3. S. Billonnet, S. Elloumi, and M-C. Plateau. Convex quadratic programming for exact solution of 0-1 quadratic programs. Technical Report 856, Laboratoire CEDRIC, Institut d’Informatique d’Entreprise, Paris, FR, 2005.
4. J. Błażewicz, P. Formanowicz, and M. Kasprzak. Selected combinatorial problems of computational biology. *European Journal of Operational Research*, 161(3):585–597, 2005.
5. A. Danchin. *The Delphic Boat*. Harvard University Press, Cambridge, MA, 2002.
6. D.B. Gordon, G.K. Hom, S.L. Mayo, and N.A. Pierce. Exact rotamer optimization for protein design. *Journal of Computational Chemistry*, 24(2):232–243, 2003.
7. H.J. Greenberg, W.E. Hart, and G. Lancia. Opportunities for combinatorial optimization in computational biology. *INFORMS Journal on Computing*, 16(3):211–231, 2004.
8. P.L. Hammer and A.A. Rubin. Some remarks on quadratic programming with 0-1 variables. *R.I.R.O.*, 4:67–79, 1970.
9. W.E. Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing (STOC)*, pages 157–168, New York, NY, 1995. ACM Special Interest Group on Algorithms and Computation Theory.
10. A. Holder, editor. *Mathematical Programming Glossary*. INFORMS Computing Society, <http://glossary.computing.society.informs.org>, 2006.
11. C.L. Kingsford, B. Chazelle, and M. Singh. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, 21(7):1028–1036, 2005.
12. G. Lancia, R. Carr, B. Walenz, and S. Istrail. 101 optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem. In *Proceedings of the Fifth Annual International Conference on Computational Biology*, pages 143–202, New York, NY, 2001.
13. S.B. Needleman and C.D. Wunch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, 48:444–453, 1970.
14. J.R. Riordan, J.M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, and J.L. Chou. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922):1066–1073, 1989.
15. M.S. Waterman. Efficient sequence alignment algorithms. *Journal of Theoretical Biology*, 108(3):333–337, 1984.
16. W. Xie and N.V. Sahinidis. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics*, 22(2):188–194, 2006.

On Value of Flexibility in Energy Risk Management. Concepts, Models, Solutions

Jörg Doege¹, Max Fehr², Juri Hinz³, Hans-Jakob Lüthi², and Martina Wilhelm²

¹ McKinsey and Company, Inc., Frankfurt, Germany
`doege@ifor.math.ethz.ch`

² Institute for Operations Research, ETH Zurich, Switzerland
`luethi@ifor.math.ethz.ch`
`martina.wilhelm@math.ethz.ch`
`maxfehr@ifor.math.ethz.ch`

³ RiskLab, ETH Zurich, Switzerland
`hinz@ifor.math.ethz.ch`

This research project is gratefully supported by the Swiss Innovation Promotion Agency KTI/CTI in collaboration with the industrial partner Nordostschweizerische Kraftwerke AG.

1 New Challenges from Deregulation

Since 90s power markets are being restructured worldwide and nowadays electrical energy is traded as a commodity. Therewith the question how to manage and hedge the financial risks resulting from uncertain electrical power and fuel prices is essential for market participants. There exists a rich literature on risk management in energy markets. Some noteworthy references can be downloaded from our web resources [1] and are reviewed in the cited literature. Let us first investigate the market structure and then discuss two different pricing schemes for risk management in power industries.

Market structure

As a first approximation, the market for electricity is not fundamentally different from any other markets. The price and the quantity of produced and consumed electricity are determined by generators costs and consumers willingness to pay.

A typical agreement traded at electricity markets yields power at constant intensity within a pre-defined time interval. However, electricity trading differs from the usual commodity trading since depending on the maturity time of the supply contract, different market players enter the transactions

- on the *long-term scale* (years to days to delivery), all agents (financial investors, suppliers, retailers, consumers) participate in contract trading
- on the *middle-term scale* (one day to delivery), at least one contract party is involved in physical consumption since positions at the day-ahead market imply physical energy delivery

- on the *short-term scale* (delivery within current day), intraday trading is effected by few agents who are able to adjust their production and demand on hourly basis

The crucial implication from this heterogeneity of the market and its participants is a large diversity with respect to

- risk attitudes (which are obviously different for producers, consumers, banks, and hedge funds)
- exposure to price risk (by their business, consumers are short, suppliers are long in electricity prices, whereas financial actors consider energy-related investments as a welcome opportunity to diversify their portfolios, or to benefit from alleged price inefficiency)
- strategies to manage the own exposure (suppliers and consumers are able to adjust their production/demand depending on market conditions, whereas financial players apply advanced knowledge in pricing and hedging of financial contracts.)
- information asymmetry (obviously agents involved in production, consumption and transmission of electrical energy benefit from insider information, whereas financial players are better skilled in econometrical techniques)

The above market heterogeneity creates an insisting need for *risk exchange*, which is effected by financial instruments. In electricity markets, we observe

- a significant diversity of contracts, explained by differences in risk profiles of market players
- exotic character of instruments, traced to replication requirement by physical assets

Implications for risk management As an outcome of the deregulation, new challenges are to be faced in energy risk management with focus on the following issues:

- competition is displaced to an intellectually higher level, where mathematically involving methodologies become more and more relevant, which changes traditional business lines
- decision makers should be aware of this development to accrue knowledge form actuarial and financial mathematics
- firm management has to correctly place the own business in the market. In particular, by the choice of a risk philosophy best suited to the business venture and by a consequent implementation of the desired policy, to correctly position the enterprise in the market.

The methodologies to meet those challenges are

1. to realize that the valuation of flexibility is one of the key issues in valuation of all energy-related assets
2. to distinguish the concepts of the individual and global viewpoint on contract valuation
3. to develop own benchmarks for pricing of flexibility
4. to develop and to implement risk-neutral models for energy-related financial contracts
5. to compare the individual and the global concepts considering both as limiting cases, with the realistic situations in between
6. to derive own decisions from this comparison

2 Implementation of Pricing Schemes

In this section, we substantiate the above lines of actions. For the sake of concreteness, we concentrate on valuation of a production capacity. To start up, we consider three typical questions which exemplarily illustrate that pricing of assets, on an abstract level, naturally reduces to the problem of *flexibility valuation*.

Example 1 An aluminum producer holds long-term supply contract. He observes that in the intraday trading, short-term electricity prices easily rise to above 250 EURO. To benefit from price spikes, he decides to install a technology to interrupt the aluminum production. By doing so, he confronts the problem whether the investment in interruptible production technology will be finally rewarded.

Example 2 The energy supplier observes that its customer (aluminum producer) achieves lucrative returns from interruptible metal production. Apparently, splitting off a constant supply contract into a delivery-on-demand and an interruptible-on-call part helps to increase its value. The energy supplier utilizes this realization by offering its customers the flexible part as a swing option and the non-flexible part as an interruptible supply contract. Here the question appears how to determine the fair price for both parts and how to justify the pricing methodology and to explain the price to the customers.

Example 3 An electricity generator owns a hydro power plant which is able to start up the production within minutes, workable to skim electricity price spikes. However, the basin is limited, hence the optimal production schedule turns out to be a non-trivial optimal control problem involving unknown statistical properties of the electricity price process and the water inflow dynamics. On the other hand, the generator realizes that a remarkable high and riskless revenue can be obtained from production-on-demand for a consumer. In other words, the supplier tends to sell the flexible hydro-electric production as a swing contract. Again, the fair price is to be determined.

In the current report we elaborate on two different pricing schemes. Both approaches are considered as limiting cases of the present situation in deregulated electricity industry. Thus, a comparison and combination of both seems to be the most reliable vehicle to establish solid asset valuation methodologies.

The above examples clarify that the optionality embedded in an agreement is well-appreciated and the market automatically assigns a price to this good. To clarify the theoretical background behind the pricing, we utilize results from modern asset pricing theory adapted to the special context of the electricity market. Below, we consider two distinct approaches, each fully justified in an idealistic situation. The reality of electricity business does not perfectly match any of the limiting cases. However, assessing a concrete case and comparing asset prices obtained in idealistic situations will help to work out the correct pricing decision.

2.1 Individual Pricing

Assumption:

There are few investors, no financial instruments are traded. Two agents exchange

the instrument at a certain price, both improve their performance by the transaction.

Important:

Who is trading (seller's and buyer's risk aversions, their portfolios, and strategies).
What is traded (contract size, policies embedded in the contract).

Ingredients:

Denote the terminal payoffs of the seller and buyer by X_S, X_B which are modeled by random variables on an appropriate probability space. Moreover, we assume that the seller is an energy producer and possesses the opportunity to change the payoff by following a policy π_S , giving the portfolio payoff $X_S^{\pi_S}$. The set of all admissible policies is denoted by $\mathcal{A}_S$. Note that a typical portfolio consists of long-term financial contracts, supply liabilities, and a production portfolio. Thus a policy π_S may be fairly complex, describing self-financed contract trading and production schedule which responds to technical constraints and environmental restrictions. Further, consider the financial contract being exchanged between the seller and the buyer, with payoff C^π depending on the optionality embedded in the contract. This optionality is securitized by the argument that the contract owner may choose any policy π from the set $\mathcal{A}_C$ of admissible policies to exercise the contract which yields C^π , whereas the seller is obliged to pay the same amount C^π . Next, we consider the risk aversions of the agents. Denote by ρ_S, ρ_B the risk measures describing sellers and buyers attitudes respectively. More precisely, a risk measure ρ is a real-valued functional defined on a set of random variables which is

- convex
- decreasing monotone
- translation invariant in the sense that

$$\rho(X + m) = \rho(X) - m \quad \text{for } m \in]-\infty, \infty[$$

The most prominent examples for risk measures are the *entropic measure*

$$\rho : X \mapsto \gamma \ln E(e^{-\frac{1}{\gamma} X}) \quad \text{with risk coefficient } \gamma \in]0, \infty[$$

and the *conditional value at risk*

$$X \mapsto E(-X | -X \geq \text{VaR}_\alpha(X))$$

where the so-called value-at-risk $\text{VaR}_\alpha(X)$ is defined by

$$\text{VaR}_\alpha(X) = \inf\{m \in \mathbb{R} : P(m + X \leq 0) \leq \alpha\}$$

with $0 < \alpha < 1$, for instance with $\alpha = 0.05$.

Remark A position which insures an uncertain income Z is interpreted as acceptable if $\rho(Z) \leq 0$. In this context, the cash amount m needed to compensate for taking the risk X is nothing but $\rho(X)$, due to

$$\rho(\underbrace{X + \rho(X)}_{=Z}) = 0$$

Efficiency Frontier

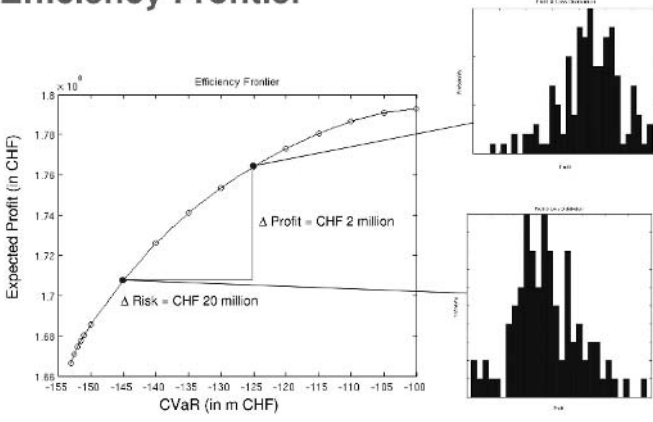


Fig. 1. Change in profit/loss distribution by optimal dispatch.

Price formation mechanism

The price c paid for the instrument $(C^\pi)_{\pi \in \mathcal{A}_C}$ is acceptable for the seller if it is sufficient to reduce its risk, regardless the strategy of the buyer:

$$\inf_{\pi_S \in \mathcal{A}_S} \rho_S(X_S^{\pi_S} - (C^\pi - c)) \leq \inf_{\pi_S \in \mathcal{A}_S} \rho_S(X_S^{\pi_S}) \quad \text{for each } \pi \in \mathcal{A}_C. \quad (1)$$

On the other hand, the price c is acceptable for the buyer, if for the best possible exercise policy, it reduces buyers own risk

$$\inf_{\pi \in \mathcal{A}_C} \rho_B(X_B + (C^\pi - c)) \leq \rho_B(X_B).$$

According to this perception, the prices at which the transaction occurs must be acceptable for both, the seller and the buyer. Let us emphasize that from sellers perspective, it is important to estimate an acceptable price from below. Thus, the seller individually has to determine prices which satisfy (1), we rewrite this inequality as

$$\inf_{\pi_S} \rho_S(X_S^{\pi_S} - C^\pi) - \inf_{\pi_S} \rho_S(X_S^{\pi_S}) \leq c \quad \text{for all } \pi \in \mathcal{A}_C.$$

Having sold the instrument, the risk increase in sellers portfolio has to be compensated by the cash amount realized from this transaction. Since the strategy $\pi \in \mathcal{A}_C$ of the counterparty is not a priori known, the precise estimation from sellers price from below

$$\inf_{\pi \in \mathcal{A}_C} \left(\inf_{\pi_S} \rho_S(X_S^{\pi_S} - C^\pi) - \inf_{\pi_S} \rho_S(X_S^{\pi_S}) \right)$$

can be replaced by a rough estimation

$$\inf_{\tilde{\pi}_S \in \tilde{\mathcal{A}}_S} \rho_S(\tilde{X}^{\tilde{\pi}_S}) - \inf_{\pi_S \in \mathcal{A}_S} \rho_S(X_S^{\pi_S}) \quad (2)$$

where $\tilde{\pi}_S$, $\tilde{\mathcal{A}}_S$ and $\tilde{X}^{\tilde{\pi}_S}$ stand for policy, set of admissible policies and the portfolio revenue in the case that the seller isolates a physical asset within the own portfolio

which is then reserved for the replication of the instrument. For instance, in the case of the virtual production capacity, $\tilde{X}_S^{\tilde{\pi}_S}$ stands for sellers portfolio not including the corresponding hydro-electric plant and the policy $\tilde{\pi}$ denotes a dispatch schedule not involving this plant.

Remark In the real-world applications, the estimation (2) requires a precise description of producers strategies which includes the specification of

- generation portfolio
- supply commitment
- long-term financial positions
- transmission constraints

The work [2] elaborates on this issues in detail. For generation portfolio modeling, the production dispatch is to optimize under technical restrictions. The corresponding techniques are presented in [9] and [3]. Figure 1 illustrates a typical risk reduction by the optimal dispatch. Thereby, the calculation of the corresponding strategy is of distinct interest, the Figure 2 shows an example therefore. Further, to model sup-

Optimal Policy to Hedge Uncertain Demand and Price

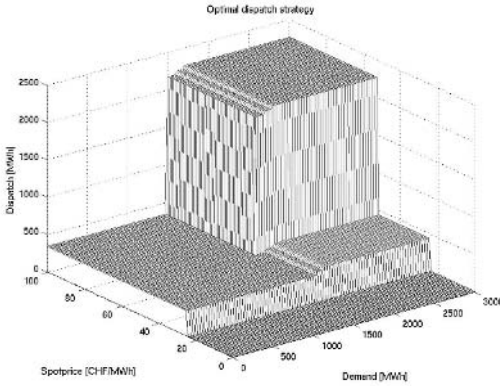


Fig. 2. Optimal dispatch strategy.

ply liabilities, electricity demand dynamics turns out to be essential. Since the time series of grind load exhibits a complex dynamics, its modeling requires advanced statistical methodologies.

2.2 Global Pricing

This pricing approach is followed in [8], [4] and [6].

Assumption:

There are many investors, diverse financial contracts are traded.

Important

Which assets are listed? (futures, forwards, virtual storages, options). Which strategies are available?

Ingredients

We start from the assumption that the financial risk hedging can be effected by appropriate set of financial instruments. Let us first introduce some terminology

- **long** position in a financial security means that the investor buys the contract and consequently gains from a price increase (and loses when price decreases). Agents holding long positions are called contract owners.
- **short** position is taken selling the security to benefit from price decrease (it causes losses when the price increases). Agents holding short positions are called contract writers.
- **derivatives** are financial agreements which ensure a certain payoff depending on market conditions, for instance on the price development of another financial contract, which is, in this context, referred to as the *underlying*.
- **no-arbitrage pricing** as the name reveals, this approach is based on the notion of the so-called *arbitrage* which stands for the opportunity of a riskless profit. Formally speaking, arbitrage is a trading strategy which produces no loss, and with a positive probability yields a gain. According to the standard perception, realistic models for asset price evolutions have to rule out any arbitrage opportunity.

We restrict ourselves to the most continent case where futures contracts are considered as prime underlyings. The price $E_t(\tau)$ at time t of a future with maturity date $\tau \in [0, T]$ is settled at a discrete number of times (for instance daily) prior to maturity. At each of those times, future price is determined by the following features:

- at expiry date $t = \tau$, the futures price $E_\tau(\tau)$ equals to the spot price E_τ of the commodity. That is, the clearing house has to ensure that the commodity can indeed be delivered immediately at τ with price $E_\tau(\tau) = E_\tau$.
- at any settlement time t_k agents holding long position in the future contract for one period (say from t_{k-1} to t_k) receive the payment $E_{t_k}(\tau) - E_{t_{k-1}}(\tau)$, whereas those holding short positions have to pay the same amount.
- the futures price $E_t(\tau)$ at the settlement time t is determined such that taking arbitrary (long, short) contract position requires no payment.

Price modeling approach

Here, leading assets are futures, say with maturities $\tau_1, \dots, \tau_N \in [0, T]$. The no-arbitrage modeling requires to introduce on the filtered probability space $(\Omega, \mathcal{F}, P, (F_t)_{t \in T})$ a selection $(E_t(\tau_i))_{t \in [0, \tau_i]}$ $i = 1, \dots, N$ of processes, each of them describes future maturing at τ_i and follows a martingale with respect to a probability measure

Q^E equivalent to P . This construction automatically excludes arbitrage for futures trading strategies. However, realistic commodity price models have to reflect physical commodity management.

We suggest an axiomatic approach (see [8], in the form below from [6]), in which commodity market model is realized on $(\Omega, F, P, (F_t)_{t \in T})$ by continuous, positive-valued adapted processes

$$[(E_t)_{t \in [0, T]}, \quad (E_t(\tau_i))_{t \in [0, \tau_i]}, \quad i = 1, \dots, N]$$

where $(E_t)_{t \in [0, T]}$ stands for the spot price process and $(E_t(\tau_i))_{t \in [0, \tau_i]}$ denote prices of futures maturing at $0 < \tau_1 < \dots < \tau_N \leq T$. Furthermore,

C1) There is no arbitrage for trading futures in the sense that there exists a measure Q^E equivalent to P such that

$$(E_t(\tau_i))_{t \in [0, \tau_i]} \text{ are } Q^E\text{-martingales for } i = 1, \dots, N$$

C2) Initial futures curve is respected

$$E_0(\tau_i) = E_0^*(\tau_i) \quad \text{for } i = 1, \dots, N$$

C3) Terminal futures price matches the spot price

$$E_{\tau_i}(\tau_i) = E_{\tau_i} \quad \text{for } i = 1, \dots, N$$

C4) Storage costs are reflected in the sense that

$$E_t(\tau_{i+1})(1 - k) \leq E_t(\tau_i) \quad \text{for all } t \in [0, \tau_i], i = 1, \dots, N$$

for a fixed $k \in]-\infty, 1]$.

Let us explain why the storage cost restrictions are, to certain degree, covered by the C4. Assume that pumping water into a hydro storage and turbinating later saves 60% of energy. In an ideal world, it would mean that

$$E_t(\tau_{i+1}) \cdot 0.6 > E_t(\tau_i)$$

for $t \in [0, \tau_i]$ is impossible (hence, $k = 1 - 0.6 = 0.4$). The justification is that on the contrary, entering one long position in the future maturing at τ_i , one short position in that maturing at τ_{i+1} , pumping 1 MWh of electricity at τ_i one turbinating the corresponding water at τ_{i+1} yields a riskless profit $E_t(\tau_{i+1}) \cdot 0.6 - E_t(\tau_i) > 0$. Note that a realistic storage costs exhibit a more complicated structure, due to its stochastic nature and dependence on inventory levels. However, C4 seems to roughly reflect storage restrictions.

Let us explain how to realize commodity price models responding to the axioms C0) – C4). It turns out that a currency change [7], [8] provides a useful connection to fixed income markets. The idea is to express all futures prices in units of commodity prices just in front of delivery. In this new currency, commodity futures behave like zero bonds given by

$$p_t(\tau) := E_t(\tau)/E_t(t) \quad t \in [0, \tau], \tau \in [0, T],$$

whereas money converts to a risky asset defined as

$$N_t := 1/E_t(t) \quad t \in [0, T].$$

Using expertise from fixed-income market theory, we suggest advanced stochastic models which satisfy axioms C0) — C4) and show how to determine no-arbitrage option prices. In particular, we discuss in [8] of fair option pricing in the case of the so-called Heath-Jarrow-Morton interest rate models. The corresponding formulas are implemented as a Java applet (see Figure 3) and are publicly available.

The numerical aspects of swing type contracts under the above presented electricity price dynamics are discussed in [7], [4]. One interesting observation from these investigations is that in the case of zero production costs (virtual hydro storage) the fair contract price does not depend on spot price volatility (see [7]).

The work presented in [11] and [10] pursues a more general approach for valuing swing type of derivatives. The proposed numerical procedure is based on finite element methods and is applicable for any continuous electricity price process. That is, the proposed numerical scheme is flexible with respect to different spot price models and different payoff functions. Moreover, the results are highly accurate and allow to interpret the influence of spot price models on swing option prices and exercise boundaries. Especially the computation of the exercise boundary is beneficial as it brings some light into the crucial question of how to exercise swing options optimally. Now, we turn to the value of flexibility addressed in the examples at the beginning

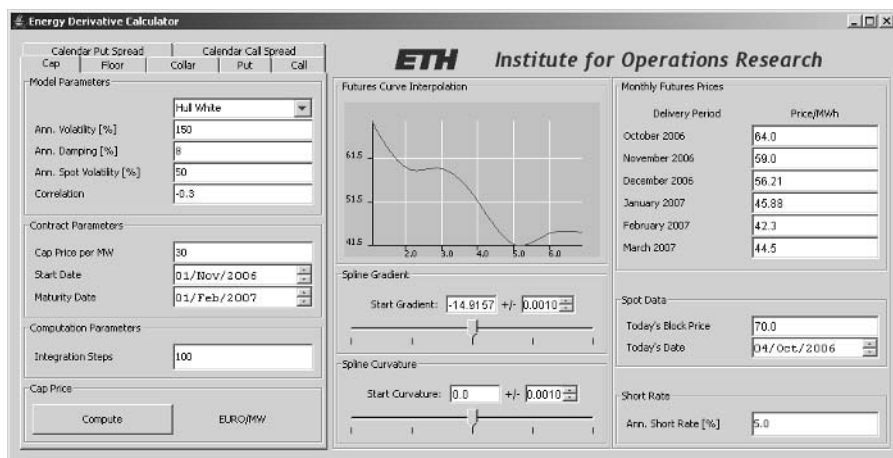


Fig. 3. Calculator for valuation of commodity options available at <http://www.ifor.math.ethz.ch/staff/hinz/CrossCommodityOptions>

of the section 2. A common problem in electricity business is to estimate the sensitivity of the hydro storage value with respect to electrical power of the turbine. The qualitative dependence is obvious: the higher the electrical power, the more flexible and thus more valuable the contract. However, only a reliable quantitative estimate can answer the question, if by upgrading the plant (say, through a costly installation of additional turbines), the producer is finally rewarded by the market. The Figure

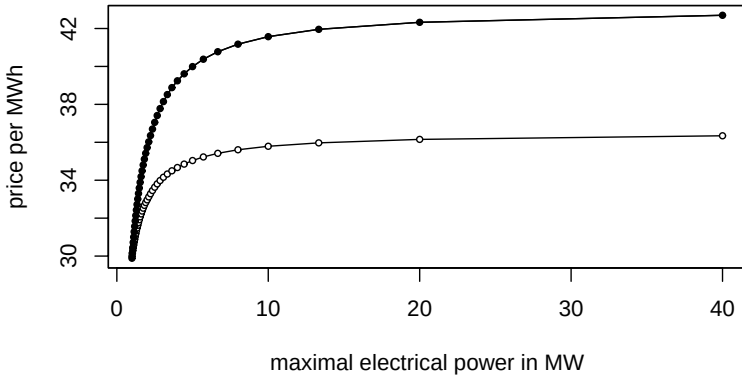


Fig. 4. The dependence of energy price on the maximal electrical power, calculated for different forward rate volatilities in the Ho-Lee model. Lines with filled circles indicate high forward rate volatilities, while lines with empty circles indicate low forward rate volatilities.

4 illustrates the dependence of the energy price within the swing contract depending on the maximal electrical production power (for details, see [7]).

3 Conclusion

In energy production, consumption and trading, all active market participants, even those avoiding any kind of speculation, are naturally exposed to several risk sources. Thereby, the price risk is substantial. Moreover, different actors (producers, consumers, financial investors) hold, according to their business, different positions and follow distinct risk aversion philosophies. This inhomogeneity makes risks exchange by financial instruments economically meaningful and individually advantageous. The theoretical foundations underlying the valuation of energy related contracts are important from practical viewpoint, however more essential is the insight that the true understanding of logical principles behind risk exchange, price of flexibility, fair value of a financial contract and its hedging helps to identify and to control the enterprise risk. As many examples of improper risk management in industry have shown, this competence turns out to be a vital functionality to survive, to position oneself and to succeed in tomorrows business.

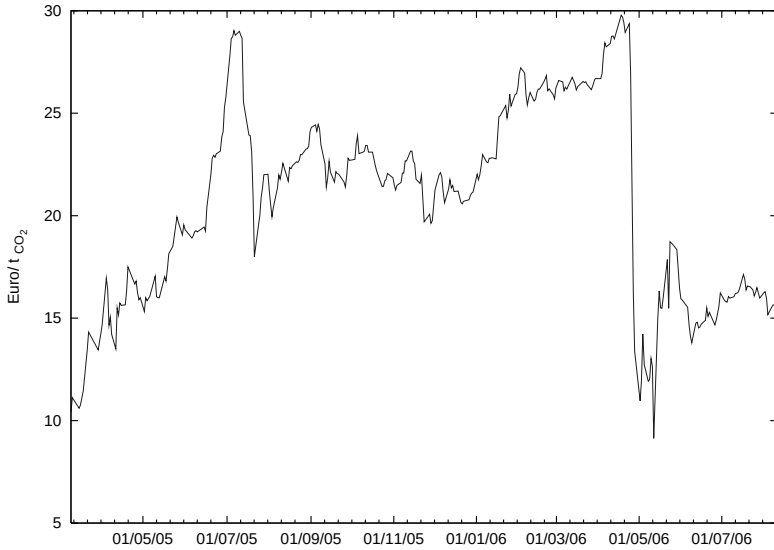


Fig. 5. Carbon allowance spot price, listed at the European Energy Exchange EEX. The recent price drop occurred while carbon emission data became public showing that the overall market position is long

4 Outlook

In electricity industry, risk management will become more demanding in the next future. Imposed by the emission reduction targets from Kyoto Protocol, participants in energy markets have to cope with considerable complexity and high price uncertainty from upcoming trading of carbon allowances. Thus, we observe that in the European energy business, a notable attention is being paid to the new theme, the so-called carbon price risk management. Let us briefly sketch the logical principles behind this issue which we have developed in [5].

On the core of our consideration is the so-called European Emission Trading Scheme (EU ETS). This mechanism is designed by the European Community to fulfill the targets under the Kyoto Protocol and is characterized by the idea to depute the responsibility for non-compliance from the level of Kyoto members (which are countries) directly to the emission sources, which are power plants and industrial users. According to the comprehensive theoretical groundwork of many economists, environmental targets can be fulfilled in the most cost-efficient way by marketable emission credits. Thus, beginning in 2005, EU has created a market for carbon allowances by the introduction of the mandatory regulatory framework of EU ETS. Several exchanges in Europe are now committed to trading of carbon allowances. The products listed there are spot and forward contracts (see Figure 5).

In our work [5] we develop an approach to describe the mechanism of carbon price formation and show how it is quantitatively linked to price drivers, namely to the evolution of energy related commodities and to emission related factors, such as

weather and climate. Moreover, we obtain quantitative answers to regulatory questions (giving an estimate to the probability of non-compliance depending relevant parameters, and on fuel prices, as well as carbon demand uncertainty). Furthermore, we estimate the social cost caused by such regulations and elaborate on risk management problems (suggesting valuation of carbon options and developing ideas on corresponding hedging strategies)

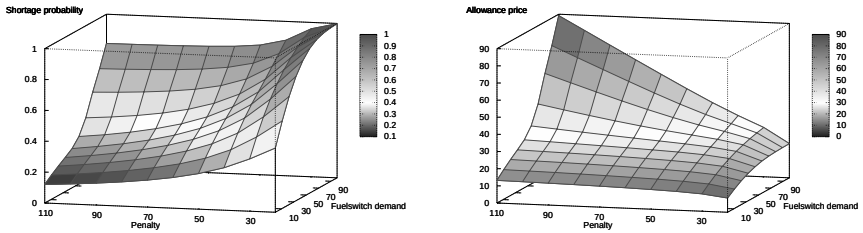


Fig. 6. The probability of non-compliance (left) and the initial allowance price (right) depending on penalty size and fuel switch demand.

References

1. http://www.ifor.math.ethz.ch/research/financial_engineering/energy.
2. J. Doege. *Valuation of flexibility for power portfolios – a dynamic risk engineering approach*. PhD thesis, ETH Zurich, Switzerland, 2006.
3. J. Doege, H.-J. Lüthi, and Ph. Schiltknecht. Risk management of power portfolios and valuation of flexibility. *OR Spectrum*, 28(2):267–287, 2006.
4. M. Fehr. Numerical methods for valuation of electricity-related contracts. Master’s thesis, ETH Zurich, Switzerland, 2005.
5. M. Fehr and J. Hinz. A quantitative approach to carbon price risk management. *Preprint*, 2006.
6. J. Hinz. Storage costs in commodity option pricing. *Preprint*, 2006.
7. J. Hinz. Valuing production capacities on flow commodities. *Mathematical Methods of Operations Research*, to appear.
8. J. Hinz and M. Wilhelm. Pricing flow commodity derivatives using fixed-income market techniques. *International Journal of Theoretical and Applied Finance*, to appear, 2006.
9. H.-J. Lüthi and Doege J. Convex risk measures for portfolio optimization and concepts of flexibility. *Mathematical Programming*, 104(2-3):541–559, 2005.
10. M. Wilhelm. *Pricing derivatives in electricity markets*. PhD thesis, ETH Zurich, Switzerland, 2007.
11. M. Wilhelm and C Winter. Finite element valuation of swing options. *Preprint*, 2006.

Bilevel Programming and Price Setting Problems

Martine Labbé

Computer Science Department, Université Libre de Bruxelles, Belgium
`mlabbe@ulb.ac.be`

Consider a general taxation model involving two levels of decision-making. The upper level (leader) imposes taxes on a specified set of goods or services while the lower level (follower) optimizes its own objective function, taking into account the taxation scheme devised by the leader. This model belongs to the class of bilevel optimization problems where both objective functions are bilinear.

In this talk, we review this class of hierarchical problems from both theoretical and algorithmic points of view.

We first introduce a general taxation model, and present some of its properties. Then, we focus on the problem of setting tolls on a specified subset of arcs of a multicommodity transportation network. In this context the leader corresponds to the profit-maximizing owner of the network, and the follower to users travelling between nodes of the network. The users are assigned to shortest paths with respect to a generalized cost equal to the sum of the actual cost of travel plus a money equivalent of travel time. Among others, we present complexity results, identify some polynomial cases and propose mixed integer linear formulations for that toll setting problem.

This talk is an overview of joint works with Luce Brotcorne, Sophie Dewez, Géraldine Heilporn, Patrice Marcotte and Gilles Savard.

Reliable Geometric Computing

Kurt Mehlhorn

Max-Planck-Institut für Informatik, Saarbrücken, Germany
`mehlhorn@mpi-inf.mpg.de`

Reliable implementation of geometric algorithms is a notoriously difficult task. Algorithms are usually designed for the Real-RAM, capable of computing with real numbers in the sense of mathematics, and for non-degenerate inputs. But, real computers are not Real-RAMs and inputs are frequently degenerate.

We start with a short collection of failures of geometric implementations. We then go on to discuss approaches to reliable geometric computation:

- realization of an efficient Real-RAM as far as it is needed for geometry (the Exact Computation Paradigm),
- algorithms with reduced arithmetic demand,
- approximation algorithms, and
- perturbation.

We also discuss the systems LEDA, CGAL, and EXACUS providing reliable implementations of geometric algorithms.

Financial Optimization

Teemu Pennanen

Helsinki School of Economics, Helsinki, Finland
`pennanen@hse.fi`

Many financial decision problems are most naturally formulated as optimization problems. This is the case, for example, in (arbitrage, utility, risk measure,...) pricing and hedging of (European, American, real,...) options, portfolio optimization and asset liability management. The optimization approach becomes even more natural in the presence of market imperfections such as transaction costs or portfolio constraints, where more traditional approaches of mathematical finance fail. Common to many financial problems, when properly formulated, is convexity with respect to the decision variables. This opens up possibilities of using numerical techniques that have been developed for large scale optimization problems.

Capital Budgeting: The Role of Cost Allocations

Ian Gow and Stefan Reichelstein

Graduate School of Business, Stanford University, U.S.A.

igow@stanford.edu

reichelstein@stanford.edu

1 Introduction

A common issue for firms is how to allocate capital resources to various investment alternatives. An extensive literature in finance has examined various aspects of capital budgeting, including capital constraints, the determination of discount rates, and alternative approaches to estimating cash flows and handling risk, such as real options techniques. In terms of organizational structure, a central feature of the capital budgeting process in large firms is that relevant information about the profitability of potential investment projects resides with one or several managers. It is generally accepted that preferences of these managers may not coincide with those of the firm's owners (the principal). Consequences of asymmetric information include strategic reporting by better-informed managers (for example, "sandbagging" or "creative optimism") and a need to measure performance *ex post*. Surveys consistently find that internal rate of return (IRR) criteria remain prevalent in capital budgeting decisions. Furthermore the use of artificially high hurdle rates suggests widespread capital rationing [15, 20].

Academic researchers and consultants have suggested that firms could create internal markets, perhaps using auction mechanisms, to solve capital budgeting problems [19, 11]. Beginning with the original work of Harris, Kriebel and Raviv (1982) and Antle and Eppen (1985), a literature in accounting and economics has examined capital budgeting with hidden information issues as the focus. This article provides a selective examination of the capital budgeting literature in accounting with a view to distilling what we now know, what we do not know and what issues seem promising for further research in this area.

2 Delegated Investment Decisions

Beginning with Rogerson's (1997) study, one branch of the recent literature on capital budgeting has focuses on a single manager who is given authority to conduct a single project that spans T periods. *Goal congruence* requires that the manager have an incentive to accept all positive NPV projects, and only those, for a broad class of managerial preferences and compensation structures. If undertaken, the

initial investment of b creates an asset with a useful life of T and operating cash flows $\tilde{c}_t = \theta \cdot x_t$ and expected net present value equal to $\text{NPV}(\theta) = \sum_{t=1}^T \gamma^t \cdot \theta \cdot x_t - b$, where $\gamma = (1 + r)^{-1}$ and r is the firm's discount rate. The parameter θ represents the private information of the manager, while the other parameters (b , $\{x_t\}$) are known to the designer of the performance measure.

The *residual income* performance measure has a long tradition in the accounting literature and has received renewed attention as part of the recent practitioner-driven movement towards economic profit plans; see, for instance Young and O'Byrne (2001). Residual income in period t is calculated as: $\pi_t = I_t - \hat{r} \cdot A_{t-1}$, where in our simplified setting accounting income is given $I_t = c_t - d_t \cdot b$ and book value evolves as $A_t = A_{t-1} - d_t b$.

Initially, book value is set equal to the investment ($A_0 = b$) and the depreciation scheme is required to be "tidy" ($\sum_t d_t = 1$). Under the *relative benefit depreciation rule*, the capital charge rate $\hat{r}$ is set equal to r and the sum of depreciation and capital charges, per dollar of initial investment, is calculated so that:

$$d_t + r \cdot AV_{t-1} = \frac{x_t \cdot b}{\sum_{i=1}^T x_i \cdot \gamma_i}.$$

The resulting performance measure π_t is goal-congruent regardless of the manager's planning horizon and intertemporal preferences, because the overall project NPV is effectively "annuitized" over the useful life of the project:

$$\pi_t = \frac{x_t}{\sum_{i=1}^T x_i \cdot \gamma^i} \text{NPV}(\theta).$$

It can be shown [16] that the combination of the residual income performance measure, the relative benefit depreciation rule and a capital charge rate equal to the owner's cost of capital is the (essentially) unique way of achieving goal congruence for a wide class of accounting-based performance measures. Clearly, the basic goal congruence scenario and its solution described here can be extended in various directions, including uncertain cash flows and the presence of multiple, temporally overlapping projects.

From a second-best contracting perspective, the natural question is whether goal-congruent performance measures can also be the basis of an optimal incentive contract in the presence of hidden action problems. Dutta and Reichelstein (2002) show that a second-best investment policy can be implemented by setting the capital charge rate equal to a *hurdle rate* which exceeds the firm's cost of capital.¹ If the depreciation charges again conform to the relative benefit rule, the residual income performance measure will result in proper matching of cash revenues and depreciation expenses so as to reflect value creation in a temporally consistent fashion. As a consequence, the choice of bonus payments in different periods has no impact on the manager's investment incentives. Put differently, accrual accounting allows the

¹ This characterization applies to a risk-neutral manager. Dutta and Reichelstein (2002) and Christensen, Feltham and Wu (2002) examine how to set the capital charge rate for a risk-averse manager who can decide on the acceptance of a risky project. In particular, these studies call into question the usual textbook recommendation of setting the capital charge rate equal to the firm's weighted-average cost of capital (WACC).

designer to separate long-term investment incentives from ongoing, short-term effort incentives.

The delegated investment problem described here has been extended in different directions. For instance, Lambert (2001) and Baldenius (2003) allow for the possibility that the manager is not intrinsically indifferent about the project but rather derives “empire benefits” from its adoption. Friedl (2005), Dutta and Reichelstein (2005) and Pfeiffer and Schneider (2006) consider sequential investment decisions with abandonment or growth options. Bareket and Mohnen (2006) examine a one-shot project selection in which the agent must choose among mutually exclusive projects.

3 Interactive Capital Budgeting Mechanisms

The selection of new investment projects frequently requires coordination among multiple business units (divisions) either because the parties are competing for scarce investment funds or because a proposed investment project generates economic benefits for multiple divisions. Suppose a firm can acquire an asset that will benefit $n > 1$ divisions. If the asset is acquired, Division i 's expected value of future cash flows is $PV_i(\theta_i) = \sum_{t=1}^T \gamma^t \cdot x_{it} \cdot \theta_i$, while the aggregate NPV is

$$NPV(\theta) = \sum_{i=1}^n PV_i(\theta_i) - b.$$

The project IRR is defined as the scalar $r^o(\theta)$ that solves

$$\sum_{i=1}^n \sum_{t=1}^T (1 + r^o(\theta))^{-t} \cdot x_{it} \cdot \theta_i = b.$$

Given the information parameters $(\theta_1, \dots, \theta_n)$, the decision rule is to invest if and only if, for all i , θ_i exceeds a critical value $\theta_i^*(\theta_{-i})$ given by $\theta_i^*(\theta_{-i}) = PV_i(\theta_i^*(\theta_{-i}) + \sum_{j \neq i} PV_j(\theta_j) = b$.

Baldenius, Dutta and Reichelstein (2006) study capital budgeting mechanisms for shared assets which require each division to report its individual profitability parameters θ_i . As part of the mechanism, the firm's central office commits to the following:

- An investment decision rule $I(\theta) \in \{0, 1\}$
- Asset shares $\lambda_i(\theta)$ satisfying $\sum_{i=1}^n \lambda_i(\theta) = 1$
- A capital charge rate $\hat{r}(\theta)$
- Depreciation schedules: $d_i(\theta)$ which satisfy $\sum_{t=1}^T d_{it}(\theta) = 1$. Thus, each division i 's asset balance at the end of period t equals

$$A_{it} = \left(1 - \sum_{\tau=1}^t d_{i\tau} \right) \cdot \lambda_i \cdot b \cdot I.$$

Restricting attention to the residual income performance measure, the above class of mechanisms imposes a “no-play-no-pay” condition requiring that a division only be charged if the project in question is undertaken. Baldenius et al. (2006) search

for *satisfactory* mechanisms which (i) are strongly goal congruent in the sense that agents have a dominant strategy incentive to report truthfully regardless of their intertemporal preferences and (ii) the project is undertaken whenever the aggregate NPV is positive. It turns out that the following Pay-the-Minimum-Necessary (PMN) mechanism is the **unique** satisfactory mechanism:

- $I(\theta) = 1$ if $\theta_i \geq \theta_i^*(\theta) - i$ for all i .
- The hurdle rate is set equal to the IRR at the critical values: $\hat{r} = r^*(\theta) \equiv r^o(\theta_1^*(\theta_{-1}), \dots, \theta_n^*(\theta_{-n}))$
- The asset shares are set in proportion to the critical values ²

$$\lambda_i(\theta) = \frac{\theta_i^*(\theta_{-i})}{\sum_{j=1}^n \theta_j^*(\theta_{-j})}$$

- Depreciation is calculated according to the relative benefit rule, relative to the hurdle rate $r^*(\theta)$.

The uniqueness of the PMN mechanism is partly due to the fact that the dominant strategy requirement implies a Groves mechanism (see, for example, Green and Laffont (1979)). The “no-play-no-pay” condition pins down the class of feasible transfer payments and the goal congruence requirement further necessitates that the transfer payments are properly allocated across time periods.

By construction, the PMN mechanism is *nominally* balanced in the sense that total depreciation charges across time periods and divisions equals the initial investment cost. Yet, the central office effectively subsidizes the investment decision since the capital charge rate, $r^*(\theta)$ is below the firm’s cost of capital, r . To see this, it suffices to note that:

$$r = r^o(\theta_{-i}, \theta_i^*(\theta_{-i})) \geq r^o(\theta_1^*(\theta_{-1}), \dots, \theta_n^*(\theta_{-n})) = r^*(\theta),$$

where the inequality follows from the definitions of the critical types and the definition of the IRR, r^o . If hidden action problems with attendant agency costs are added to the model, the optimal capital charge (hurdle) rate will increase relative to the goal congruence setting, just as in the single division scenario. With sufficiently high agency costs, the second-best hurdle rate may exceed r . Thus the planning and the control problem pull the optimal hurdle rate in opposite directions with the ultimate outcome dependent on the relative magnitude of the two factors.

In contrast to the shared-asset scenario described above, it is natural to consider settings in which the firm must choose one project from those submitted by each of n divisions. As the divisions now compete against each other, the results are to some extent “dual” to the ones in the shared asset scenario. Focusing on a class of capital budgeting mechanisms similar to the one described above, Baldenius et al. (2006) identify the Competitive Hurdle Rate (CHR) mechanism as the uniquely satisfactory mechanism in this setting. In particular, divisional managers are asked to “bid” the IRR of their projects, the highest bidder is awarded the project (provided his/her IRR exceeds r) and the winning division faces a capital charge rate equal to the IRR of the second-highest bidder.³ In contrast to the shared asset coordination problem,

² This characterization only applies to the special case of ex-ante identical divisions; see Baldenius et al. (2006).

³ This simple second-price auction characterization applies when the divisional projects are ex-ante identical.

the resulting competitive hurdle rate will now exceed r . On the other hand, when moral hazard problems are added to the model, a uniform increase in agency costs by all divisions will generally not raise the hurdle rate further, since it is the relative agency costs that matter when the divisional projects compete against each other. In contrast, the sum of the individual agency costs determine the hurdle rate for shared assets.

4 Capacity Investments and Capacity Management

In the preceding settings, a project, once undertaken, does not require any further managerial decisions. A natural extension of these models therefore is one where the investment decisions pertain to capacity investments, yet the firm will decide on capacity utilization at a later time. Rogerson (2006) develops a model in which an asset acquired at date t creates capacity to produce k widgets in each of the next T years (the asset is worthless thereafter). As such, multiple generations of assets create production capacity K_t at date t . The contribution margin, $R_t(K_t)$ associated with this capacity is not known to the firm's central office, but this office does know that the optimal capacity sequence if increasing ($K_{t+1}^* > K_t^*$ for all t).

Rogerson (2006) shows that given an infinite planning horizon, goal congruence can be attained if (i) the capital charge rate is set to r , the firm's discount rate, and (ii) assets are depreciated using the annuity method. Thus the depreciation charges are not calibrated to the *relative* magnitude of expected future cash benefits, $R_t(\cdot)$, as advocated in the above models. The reason is that the sequence of overlapping investment decisions creates an effective intertemporal separation in the capacity decisions, yet intertemporal cost allocation is still essential in order for managers to internalize the *marginal* cost of capacity.

This result has powerful implications for internal pricing. Suppose a downstream divisions relies on widgets, supplied by an upstream division, as a production input. If widgets are priced at full cost, comprising any applicable variable production costs and the *unit (marginal)* cost of capacity required to produce the widget, goal congruence is obtained in that (i) the downstream division will demand the efficient quantity of widgets and (ii) the upstream division will have an incentive to invest optimally. However, this efficiency result essentially requires a "clockwork" environment in which the supplying division is never saddled with excess capacity or capacity shortages due to demand shocks in any given period. The possibility of such fluctuations has, of course, been the focus of recent studies on capacity trading mechanisms in supply chains [14].

5 Concluding Remarks

There has been significant progress in recent years in understanding the role of intertemporal cost charges, specifically depreciation and capital charges, in the capital budgeting process. Cost allocations across time periods and business units are essential in creating time-consistent performance measures that compare initial reports to subsequently realized outcomes. But major challenges remain in making the existing

models more complete on several fronts, including (i) richer settings of capacity investments and subsequent capacity utilization, (ii) projects with sequential decision stages and (iii) the coordination of risk exposure across the divisions of a firm.

References

1. Antle, R. and Eppen, G. D. 1985, Capital rationing and organizational slack in capital budgeting, *Management Science* **31**(2), 163–174.
2. Baldenius, T. 2003, Delegated investment decisions and private benefits of control, *Accounting Review* **78**(4), 909–930.
3. Baldenius, T., Dutta, S. and Reichelstein, S. 2006, Cost allocation for capital budgeting decisions. Working paper.
4. Bareket, M. and Mohnen, A. 2006, Performance measurement for investment decisions under capital constraints. Forthcoming in *Review of Accounting Studies*.
5. Christensen, P. O., Feltham, G. A. and Wu, M. G. H. 2002, "Cost of capital" in residual income for performance evaluation, *Accounting Review* **77**(1), 1–23.
6. Dutta, S. and Reichelstein, S. 2002, Controlling investment decisions: Depreciation- and capital charges, *Review of Accounting Studies* **7**(2 - 3), 253–281.
7. Dutta, S. and Reichelstein, S. 2005, Stock price, earnings, and book value in managerial performance measures, *The Accounting Review* **80**(4), 1069–1100.
8. Friedl, G. 2005, Incentive properties of residual income when there is an option to wait, *Schmalenbach Business Review* **57**, 3–21.
9. Green, J. and Laffont, J.-J. 1979, *Incentives in Public Decision-Making*, North-Holland, Amsterdam.
10. Harris, M., Kriebel, C. H. and Raviv, A. 1982, Asymmetric information, incentives and intrafirm resource allocation, *Management Science* **28**(6), 604–620.
11. Hodak, M. 1997, The end of cost allocations as we know them, *Journal of Applied Corporate Finance* **10**(3), 117–124.
12. Lambert, R. A. 2001, Contract theory and accounting, *Journal of Accounting and Economics* **32**, 3–87.
13. Pfeiffer, T. and Schneider, G. 2006, Residual income-based compensation plans for controlling investment decisions under sequential private information. Working paper.
14. Plambeck, E. L. and Taylor, T. A. 2005, Sell the plant? the impact of contract manufacturing on innovation, capacity, and profitability, *Management Science* **51**, 133–150.
15. Poterba, J. M. and Summers, L. H. 1995, A CEO survey of U.S. companies' time horizons and hurdle rates, *Sloan Management Review* **37**(1), 43–53.
16. Reichelstein, S. 1997, Investment decisions and managerial performance evaluation, *Review of Accounting Studies* **2**, 157–180.
17. Rogerson, W. P. 1997, Intertemporal cost allocation and managerial investment incentives: A theory explaining the use of economic value added as a performance measure, *Journal of Political Economy* **105**(4), 770–795.
18. Rogerson, W. P. 2006, Intertemporal cost allocation and investment decisions. Working paper.

19. Stein, J. C. (1997) Internal capital markets and the competition for corporate resources, *Journal of Finance* **52**(1), 111–33.
20. Taggart, R. A. 1987, Allocating capital among a firm's divisions: Hurdle rates versus budgets, *Journal of Financial Research* **10**(3).
21. Young, S. D. and O'Byrne, S. F. 2001, *EVA and Value-Based Management: A Practical Guide*, McGraw-Hill, New York, NY.

An Overview on the Split Delivery Vehicle Routing Problem

Claudia Archetti and Maria Grazia Speranza

Department of Quantitative Methods, University of Brescia, Italy
archetti@eco.unibs.it
speranza@eco.unibs.it

Summary. In the classical Vehicle Routing Problem (VRP) a fleet of capacitated vehicles is available to serve a set of customers with known demand. Each customer is required to be visited by exactly one vehicle and the objective is to minimize the total distance traveled. In the Split Delivery Vehicle Routing Problem (SDVRP) the restriction that each customer has to be visited exactly once is removed, i.e., split deliveries are allowed. In this paper we present a survey of the state-of-the-art on this important problem.

1 Introduction

We consider the *Split Delivery Vehicle Routing Problem* (SDVRP) where a fleet of capacitated homogeneous vehicles has to serve a set of customers. Each customer can be visited more than once, contrary to what is usually assumed in the classical *Vehicle Routing Problem* (VRP), and the demand of each customer may be greater than the capacity of the vehicles. There is a single depot for the vehicles and each vehicle has to start and end its tour at the depot. The problem consists in finding a set of vehicle routes that serve all the customers such that the sum of the quantities delivered in each tour does not exceed the capacity of a vehicle and the total distance traveled is minimized.

The SDVRP was introduced in the literature only a few years ago by Dror and Trudeau ([13] and [14]) who motivate the study of the SDVRP by showing that there can be savings generated by allowing split deliveries. Archetti, Savelsbergh and Speranza [3] study the maximum possible savings obtained by allowing split deliveries, while in [4] the same authors present a computational study to show how the savings depend on the characteristics of the instance. Valid inequalities for the SDVRP are described in [12]. In [9] a lower bound is proposed for the SDVRP where the demand of each customer is lower than the capacity of the vehicles and the quantity delivered by a vehicle when visiting a customer is an integer number. In [2] the authors analyze the computational complexity of the SDVRP and the case of small capacity of the vehicles.

Heuristic algorithms for the SDVRP can be found in [13] and [14], where a local search algorithm is proposed, in [1] for a tabu search and in [5] for an optimization-

based heuristic. In [15] the authors present a mathematical formulation and a heuristic algorithm for the SDVRP with grid network distances and time windows constraints.

Real applications of the problem can be found in [18] where the authors consider the problem of managing a fleet of trucks for distributing feed in a large livestock ranch which is formulated as a split delivery capacitated rural postman problem with time windows. Several heuristics are proposed to solve the problem which compare favorably with the working practices on the ranch. Sierksma and Tijssen [19] consider the problem of determining the flight schedule for helicopters to off-shore platforms for exchanging crew people employed on these platforms. The problem is formulated as an SDVRP and several heuristics are proposed. In [6] Archetti and Speranza consider a waste collection problem where vehicles have a small capacity and customers can have demands larger than the capacity. A number of constraints are considered like time windows, different types of wastes, priorities among customers and different types of vehicles. They propose a heuristic algorithm that beats the solution implemented by the company which carries out the service. A similar problem is analyzed in [8], where it is called the Rollon-Rolloff Vehicle Routing Problem (RRVRP), and in [11].

A detailed survey of the state-of-the-art on the split delivery vehicle routing problem can be found in [7].

The paper is organized as follows. In Section 2 we provide the problem description and we present computational complexity results and some properties of the problem. In Section 3 we analyze the savings with respect to the VRP and a simple heuristic for the VRP and the SDVRP. In Section 4 we present the heuristic algorithms proposed for the SDVRP and compare them.

2 The Split Delivery Vehicle Routing Problem

The SDVRP can be defined over a graph $G = (V, E)$ with vertex set $V = \{0, 1, \dots, n\}$ where 0 denotes the depot while the other vertices are the customers, and E is the edge set. The traversal cost (also called length) c_{ij} of an edge $(i, j) \in E$ is supposed to be non-negative and to satisfy the triangle inequality. An integer demand d_i is associated with each customer $i \in V - \{0\}$. An unlimited number of vehicles is available, each with a capacity $Q \in \mathbb{Z}^+$. Each vehicle must start and end its route at the depot. The demands of the customers must be satisfied, and the quantity delivered in each tour cannot exceed Q . The objective is to minimize the total distance traveled by the vehicles. A mixed integer programming formulation for the SDVRP is provided in [1].

In [2] it is shown that the SDVRP with $Q = 2$ can be solved in polynomial time, while it is NP-hard for $Q \geq 3$.

Dror and Trudeau [13] have shown an interesting property of optimal solutions to the SDVRP. To understand their result we first need the following definition.

Definition 1. Consider a set $C = \{i_1, i_2, \dots, i_k\}$ of customers and suppose that there exist k routes $r_1, \dots, r_k$, $k \geq 2$, such that r_w contains customers i_w and i_{w+1} , $w = 1, \dots, k-1$, and r_k contains customers i_1 and i_k . Such a configuration is called a *k-split cycle*.

Dror and Trudeau have shown that, if the distances satisfy the triangle inequality, then there always exists an optimal solution to the SDVRP which does not contain k -split cycles, $k \geq 2$. This implies that there exists an optimal solution to the SDVRP where no two routes have more than one customer with a split delivery in common.

A set of valid inequalities has been proposed in [12] and a lower bound for the SDVRP is proposed in [9] which is, to the best of our knowledge, the only lower bound proposed in the literature for the SDVRP.

In [16] an exact algorithm is proposed for the SDVRP with time windows. The algorithm is introduced in [17] where instances with up to 25 customers are solved. In [16] the algorithm is improved and solves almost all instances with up to 50 customers and a subset of instances with 100 customers.

3 SDVRP vs VRP

The interest in the SDVRP comes from the fact that costs can be reduced with respect to the costs of the VRP by allowing split deliveries. In this section, we discuss the amount of the saving. This is an important information in practice, because of the additional organizational difficulties deriving from the multiple visits to the same customer. In the following we will consider both the case where the demand of each customer is lower than or equal to the capacity Q and the case where the demand of a customer can be greater than Q . For this latter case there is the need to define a variant of the classical VRP since, when the demand of a customer is greater than the vehicle capacity, it has to be split and the customer has to be visited more than once. In order to distinguish the cases, in this section we define as *extended VRP* the problem where each customer is visited the minimum number of times and *extended SDVRP* the problem where this restriction is relaxed.

In [3] it has been shown that both in the basic and in the extended version, the cost of an optimal solution of the VRP can be as large as twice the cost of an optimal solution of the SDVRP.

In the same paper the performance of the following simple heuristic has been analyzed: make full truckload deliveries using out-and-back tours to customers with demand greater than the vehicle capacity until their remaining demand is less than or equal to the vehicle capacity. Then, in order to find a minimum cost set of routes serving the remaining demands of all customers, the VRP is optimally solved, obtaining a heuristic for the extended version of the VRP, or the SDVRP is optimally solved, obtaining a heuristic for the extended version of the SDVRP. For both problems, the heuristic can find a solution whose cost is as large as twice the cost of the optimal solution.

While the worst-case results discussed above are of great theoretical and also practical relevance, important additional information can be obtained from an empirical study. In [4] the results obtained from a computational study are presented that confirm that allowing split deliveries can result in substantial benefits, but also show that these substantial benefits only occur for customers demands with fairly specific characteristics. The following insights have been obtained:

1. the benefits from allowing split deliveries mainly depend on the relation between mean demand and vehicle capacity and on demand variance; there does not appear to be a dependence on customer locations;

2. the major benefit of allowing split deliveries appears to be the ability to reduce the number of delivery routes and, as a consequence, the cost of the routes (this also explains the fact that there does not appear to be a dependence on customer locations);
3. the largest benefits are obtained when the mean demand is greater than half the vehicle capacity but less than three quarters of the vehicle capacity and the demand variance is relatively small.

4 Heuristics for the SDVRP

To the best of our knowledge, three heuristic algorithms have been proposed to solve the SDVRP: a local search heuristic by Dror and Trudeau [13], a tabu search heuristic by Archetti, Hertz and Speranza [1] and an optimization-based heuristic by Archetti, Savelsbergh and Speranza [5]. We are also aware of a heuristic by Chen, Golden and Wasil [10], but since the paper did not yet appear we are unable to provide any additional information.

The algorithm by Dror and Trudeau [13] is designed only for the case where the demand of each customer is lower than the capacity of the vehicles. The heuristic is a local search algorithm that is very fast and provides good solutions.

In [1] a tabu search algorithm for the SDVRP, called SPLITABU, is presented and tested. It is a very simple algorithm, easy to implement, where there are only two parameters to be set: the length of the tabu list and the maximum number of iterations the algorithm can run without improvement of the best solution found.

The optimization heuristic proposed in [5] makes use of information provided by the tabu search presented in [1] in order to construct a set of good routes. These routes are then passed to a MIP program which determines the best ones. One of the key ideas underlying this solution approach is that the tabu search can identify parts of the solution space that are likely to contain high quality solutions. A set of edges which are likely to be included in high-quality SDVRP solutions is identified. If an edge is never, or rarely, included in the routes of solutions encountered during the tabu search, this information is interpreted as an indication that it is likely that the edge will not be included in high quality routes. Edges that are often included are promising edges. Then, the set of good routes is created from the promising edges. Finally, a MIP program is solved to identify the best routes and the quantities delivered by each route to each customer.

Computational results that compare the above heuristics can be found in [5] and in [7]. The best solutions are found by the optimization heuristic. However, slightly worse solutions can be obtained by the tabu search heuristic in a short time and good, though worse, solutions can be found almost in real time by the local search heuristic.

5 Conclusions

In this survey we have summarized what is known on the Split Delivery Vehicle Routing Problem which is a Vehicle Routing Problem where the classical single visit assumption is relaxed. Thus, each customer can be visited more than once.

The interest that this problem has raised is justified by the cost savings that can be obtained by allowing split deliveries. The problem has been introduced quite recently in the literature but in the last years an increasing number of researchers has focused on this subject. However, there is still a lot of work to do, as confirmed by the fact no exact algorithm has yet been proposed in the literature for the problem.

References

1. Archetti C, Hertz A, Speranza MG (2006) A tabu search algorithm for the split delivery vehicle routing problem. *Transportation Science* 40: 64–73
2. Archetti C, Mansini R, Speranza MG (2005) Complexity and reducibility of the skip delivery problem. *Transportation Science* 39: 182–187
3. Archetti C, Savelsbergh MWP, Speranza MG (2006) Worst-case analysis for split delivery vehicle routing problems. *Transportation Science* 40: 226–234
4. Archetti C, Savelsbergh MWP, Speranza MG, To split or not to split: That is the question. *Transportation Research E*, to appear.
5. Archetti C, Savelsbergh MWP, Speranza MG, An optimization-based heuristic for the split delivery vehicle routing problem. Submitted
6. Archetti C, Speranza MG (2004) Vehicle routing in the 1-skip collection problem. *Journal of the Operational Research Society* 55: 717–727
7. Archetti C, Speranza MG, The split delivery vehicle routing problem: A survey. Submitted
8. Ball M, Bodin L, Baldacci R, Mingozzi A (2000) The rollon-rolloff vehicle routing problem. *Transportation Science* 34: 271–288
9. Belenguer JM, Martinez MC, Mota E (2000) A lower bound for the split delivery vehicle routing problem. *Operations Research* 48: 801–810
10. Chen S, Golden B, Wasil E, The split delivery vehicle routing problem: A new heuristic and benchmark problems. Presented at ROUTE 2005
11. De Meulemeester L, Laporte G, Louveaux FV, Semet F (1997) Optimal sequencing of skip collections and deliveries. *Journal of the Operational Research Society* 48: 57–64
12. Dror M, Laporte G, Trudeau G (1994) Vehicle routing with split deliveries. *Discrete Applied Mathematics* 50: 239–254
13. Dror M, Trudeau G (1989) Savings by split delivery routing. *Transportation Science* 23: 141–145
14. Dror M, Trudeau G (1990) Split delivery routing. *Naval Research Logistics* 37: 383–402
15. Frizzell PW, Giffin JW (1995) The split delivery vehicle scheduling problem with time windows and grid network distances. *Computers and Operations Research* 22: 655–667
16. Gendreau M, Dejax P, Feillet D, Gueguen C, Vehicle routing with time windows and split deliveries. Working paper
17. Gueguen C (1999) Méthodes de résolution exacte pour les problèmes de tournées de véhicules. PhD thesis, École Centrale Paris, Paris
18. Mullaseril PA, Dror M, Leung J (1997) Split-delivery routing in livestock feed distribution. *Journal of the Operational Research Society* 48: 107–116
19. Sierksma G, Tijssen GA (1998) Routing helicopters for crew exchanges on off-shore locations. *Annals of Operations Research* 76: 261–286

Collaborative Planning - Concepts, Framework and Challenges

Hartmut Stadtler

University of Hamburg, Germany
`hartmut.stadtler@uni-hamburg.de`

SCM is concerned with the coordination of material, information and financial flows within and across legally separated organizational units. Software vendors have developed so called Advanced Planning Systems (APS) to provide decision support at different hierarchical planning levels of an intra-organizational supply chain. However, since APS are based on principles of hierarchical planning further ideas and concepts are required to coordinate activities and flows between adjacent planning levels of each partner.

We will define Collaborative Planning (CP) as the alignment of plans at the same planning level between independent organizational units having agreed on a supplier-buyer partnership. CP software modules today basically support the exchange of (demand and procurement) data plus some additional feasibility checks. According to Schneeweiss (2003, p. 5) this is only the starting point of CP.

In recent years a number of concepts have been advocated to support CP like agent technology, contract design and negotiation procedures. Based on these concepts an analysis of its assumptions and resulting decision support will yield a set of criteria for discriminating CP approaches. These criteria will give rise to a framework for CP to position existing concepts and to indicate blind spots in CP research.

Reference

Schneeweiss, C., Distributed Decision Making, Berlin et al. (Springer), 2nd ed., 2003.

Promoting ε -Efficiency in Multiple Objective Programming: Theory, Methodology, and Application

Margaret Wiecek and Alexander Engau

Clemson University, U.S.A.

wmalgor@clemson.edu

aengau@clemson.edu

A major goal in any optimization or decision-making process is to identify a single or all best solutions within a set of feasible alternatives. In multiple objective programming (MOP), while it is theoretically possible to identify the complete set of efficient solutions, finding an exact description of this set often turns out to be practically impossible or computationally too expensive. Thus many research efforts during the last thirty years have focused on concepts and procedures for the efficient set approximation.

In practice, the decision maker is frequently satisfied with suboptimal decisions provided the loss in optimality is justified by a significant gain with respect to model simplicity and/or computational benefits. The consideration of this trade-off reflects the common belief that the concept of ε -efficient solutions accounts for modeling limitations or computational inaccuracies, and therefore is tolerable rather than desirable. Consequently, the quest to formalize and deepen the understanding of mathematical concepts has motivated operations researchers to obtain theoretical results on ε -efficiency in MOP. However, solution methods purposely avoiding efficiency while seeking to guarantee ε -efficiency have not been developed.

The recognition of significance of ε -efficient solutions in engineering design optimization has motivated this study. Algebraic characterizations of ε -efficient solutions using polyhedral translated cones are presented. Two classes of methods, approximate and exact, for the generation of these solutions are proposed. It is also shown how ε -efficient solutions can reduce the high dimensionality in MOP due to a large number of criteria. An interactive method that allows the decision maker to consider pairs of criteria at a time, thereby making tradeoff evaluation the simplest possible, is presented. The applicability of the proposed methods to structural design is demonstrated.

Business Intelligence, Forecasting and
Marketing

Combining Support Vector Machines for Credit Scoring

Ralf Stecking and Klaus B. Schebesch

Department of Economics, University of Bremen, Germany

`stecking@uni-bremen.de`

`kbsbase@gmx.de`

Summary. Support vector machines (SVM) from statistical learning theory are powerful classification methods with a wide range of applications including credit scoring. The urgent need to further boost classification performance in many applications leads the machine learning community into developing SVM with multiple kernels and many other combined approaches. Owing to the huge size of the credit market, even small improvements in classification accuracy might considerably reduce effective misclassification costs experienced by banks. Under certain conditions, the combination of different models may reduce or at least stabilize the risk of misclassification. We report on combining several SVM with different kernel functions and variable credit client data sets. We present classification results produced by various combination strategies and we compare them to the results obtained earlier with more traditional single SVM credit scoring models.

1 Introduction

Classifier combination is a recurring topic since classification models are successfully applied across disciplines ranging from pattern recognition in astronomy to automatic credit client scoring in finance. Classifier combination is done in the hope of improving the out-of-sample classification performance of single *base* classifiers. As is well known by now ([1],[2]), the results of such combiners can be both better or worse than expensively crafted single models. In general, as the base models are less powerful (and much more easy to produce) their combiners tend to yield much better results. However, this advantage is decreasing with the quality of the base models (e.g. [1]). Our past credit scoring single-SVM classifiers concentrate on misclassification performance obtainable by different SVM kernels, different input variable subsets and financial operating characteristics ([3],[4],[5],[6]). In credit scoring, classifier combination using such base models may be very useful indeed, as small improvements in classification accuracy matter and as *fusing* models on different inputs may be required by practice. Hence, the paper presents in sections 2 and 3 model combinations with base models on all available inputs using single classifiers with six different kernels, and finally in section 4 SVM model combinations of base models on reduced inputs using the same kernel classifier.

Table 1. Region error w.r.t. figure 1 below. For critical/critical region V (where $|f_k(x)| < 1$ for both models) classification error is highest. For censored combination models these input region would not be predictable.

Region	II	III	IV	V	VI	VII	VIII	Total
Error in %	15.0	12.2	22.9	39.7	22.9	9.1	11.1	28.0
N	20	82	35	360	35	99	27	658

2 Censored Combination of SVM

The basic data set for our past credit scoring models is an equally distributed sample of 658 clients for a building and loan credit with a total number of 40 input variables. In order to forecast the defaulting behavior of new credit clients, a variety of SVM models were constructed, in part also for addressing special circumstances like asymmetric costs of misclassification and unbalanced class sizes in the credit client population [3]. Using a **fixed** set of 40 **input variables**, six different SVM with **varying kernel functions** are combined for classifying good and bad credit clients. Detailed information about kernels, hyperparameters and tuning can be found in [6].

Our combined approach of different SVM will be described as follows: Let $f_k(x) = \langle \Phi_k(x), w_k \rangle + b_k$ be the output of the k th SVM model for unknown pattern x , with b_k a constant, Φ_k the (usually unknown) feature map which lifts points from the input space $\mathbf{X}$ into feature space $\mathcal{F}$, hence $\Phi : \mathbf{X} \rightarrow \mathcal{F}$. The weight vector w_k is defined by $w_k = \sum_i \alpha_i y_i \Phi_k(x_i)$ with α_i the (C bounded) dual variables ($0 \leq \alpha_i \leq C$), and y_i be the binary output of input pattern x . Note also, that $\langle \Phi_k(x), \Phi_k(x_i) \rangle = K(x, x_i)$, where K is a kernel function, for example $K(x, x_i) = \exp(-s\|x - x_i\|^2)$, i.e. the well known RBF kernel with user specified kernel parameter s . In previous work [4] it was shown that SVM **output regions** can be defined in the following way: (1) if $|f_k(x)| \geq 1$, then x is called a *typical* pattern with low classification error, (2) if $|f_k(x)| < 1$, then x is a *critical* pattern with high classification error. Combining SVM models for classification we calculate sign $(\sum_k f_k(x))$ introducing the restriction, that $f_k(x) = 0$, if $|f_k(x)| < 1$, which means: SVM model k has zero contribution for its critical patterns. For illustrative purpose we combine two SVM models (RBF and second degree polynomial) and mark nine regions (see figure 1): typical/typical regions are I, III, VII, IX, critical/critical region is V and typical/critical regions are II, IV, VI, VIII. Censored classification uses only typical/typical regions (with a classification error of 10.5 %) and typical/critical regions (where critical predictions are set to zero) with a classification error of 18.8 %. For the critical/critical region V no classification is given, as the expected error within this region would be 39.7 % (see table 1). The crucial point for this combination strategy is the fairly high number of unpredictable patterns (360 out of 658) in this case. However, by enhancing the diversity and by increasing the number of SVM models used in combinations, the number of predictable patterns will also increase, as will be shown in the following section.

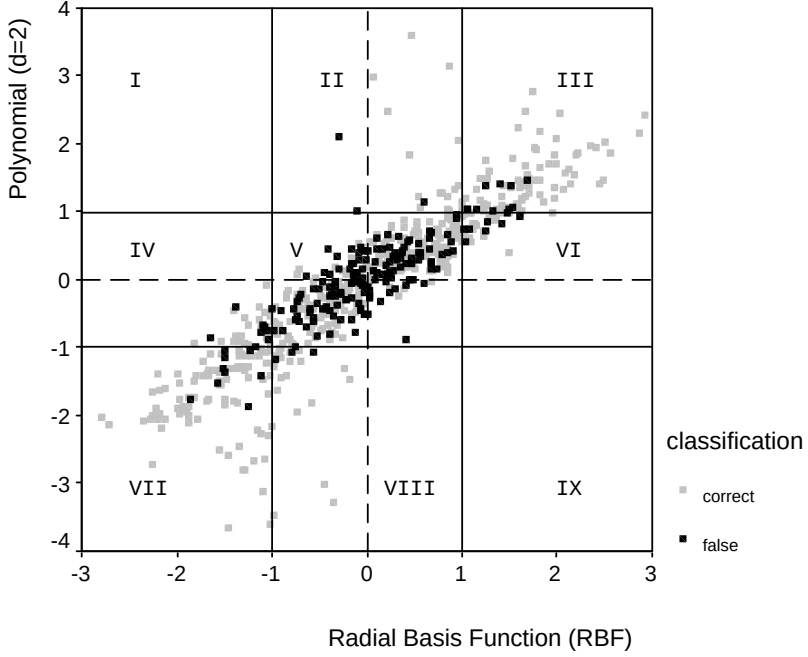


Fig. 1. Combined predictions of SVM with (i) polynomial kernel $K(x, x_i) = (\langle x, x_i \rangle + 5)^2$ and (ii) RBF kernel $K(x, x_i) = \exp(-0.05\|x - x_i\|^2)$. Black (grey) boxes represent false (correct) classified credit clients. Nine regions (I-IX) are defined s.t. the SVM output of both models (see also table 1).

3 Combining Six Models

Table 2 shows the classification results of six single SVM and two combination models after tenfold cross validation. The total error is the percentage of those clients classified incorrectly relative to all credit clients. The alpha error is the percentage of *accepted bad* relative to all bad clients and the beta error is the percentage of *rejected good* relative to all good clients. In terms of total and beta error the RBF model dominates all other single models. Coulomb has the lowest alpha error. The first combined approach *Combi Six* sums up the output values of the six single models, hence the decision function reads $\text{sign} \sum_{k=1}^6 f_k(x)$. This strategy leads only to medium performance in terms of classification error. Therefore, our second approach is introduced, the censored combination of six single models. The decision function again is $\text{sign} \sum_{k=1}^6 f_k(x)$, but here $f_k(x)$ is set to zero, if $|f_k(x)| < 1$. This combination leads to a total error of 18.9%, which is far below the best error of the six single models. However, as *Censored Combi Six* is only valid for 413 patterns, there still remain 255 unlabeled credit applicants. In order to compare the accuracy of models with different numbers of classifications, we use Pearson's χ^2 , which depends both on classification accuracy and sample size. Again, *Censored Combi Six* dominates all other models. Finally, for non standard situations (unbalanced class

Table 2. Tenfold cross validation error for six single models and two combination models. Pearson’s χ^2 is used to compare classification performance of models with different sample size.

SVM model	Alpha error	Beta error	Total error	No. of observ.	Pearson χ^2
<i>Linear</i>	29.1 %	29.9 %	29.5 %	658	110.8
<i>Sigmoid</i>	28.8 %	29.6 %	29.2 %	658	114.1
<i>Polynomial ($d = 2$)</i>	27.2 %	28.1 %	27.7 %	658	131.4
<i>Polynomial ($d = 3$)</i>	27.9 %	29.0 %	28.4 %	658	122.7
<i>RBF</i>	28.2 %	23.9 %	26.0 %	658	151.6
<i>Coulomb</i>	26.6 %	26.0 %	26.3 %	658	147.9
<i>Combi Six</i>	28.5 %	28.7 %	28.6 %	658	120.8
<i>Censored Combi Six</i>	19.3 %	18.5 %	18.9 %	413	159.8

sizes and different costs of misclassification), we suggest to use a *default class* in the combination model, which would be "good" in our case. It can be shown, that *Censored Combi Six* again leads to superior results.

4 Combination of SVM on Input Subsets

Based on previous work on input subset selection [5], we now use a SVM model population built on randomly selected features from our credit scoring data, leading to *weak* models. The model population used here consist of 1860 base models and is restricted to SVM with RBF kernels. It starts with 60 variations of five randomly chosen inputs and ends with 60 variations of 35 input variables of the original data set, respectively. This set of differently informed (or weak) base models is trained and validated by using a fast, simple and highly automated model building procedure with a minimum of SVM parameter variation. The SVM outputs of these differently sized models are used to build combiners by (a) fusing them into simple (additive) rules, and (b), by using a *supervised combiner*. We first inquire into the effects of additive SVM output combination on misclassification performance, for which the following rules are used: (1) LINCUM, which adds the base model outputs as they occur during the population built-up (finally combing 1860 models), (2) I60, which adds the outputs of models with the same number of inputs producing 31 combiners, and (3) E3, adding the outputs of the three best (elite) models from the sorted list of the leave-one-out (l-1-o) error of the 31 base model subpopulations. The experiments shown in fig.2 (left plot) indicate that these combiners consistently outperform the minimum l-1-o errors of the base models, with I60 having the best results on small input sets. Next, the inputs of the combiners E3 and I60 are used by 31 supervised combiners (also SVM models with RBF-kernels, for convenience) denoted by SVM(aE3) and SVM(aI60), respectively. These models are then subjected

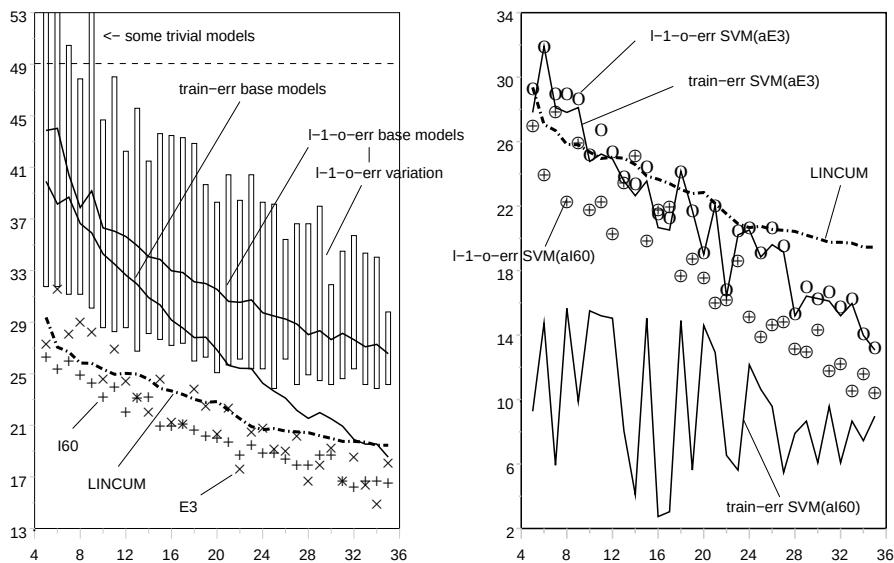


Fig. 2. Left plot: errors of base SVM classifier population using 5 to 35 random inputs and errors of linear combiners on their outputs. Right plot: errors of supervised combiners on the inputs of the linear combiners. LINCUM error is plotted again (rhs plot) to ease comparison. Errors are given in percent misclassification.

to l-1-o error validation (fig.2, right plot), which shows still further partial improvement compared to the simpler combination rules. We note that this procedure can be very useful for problems with large feature dimensions: it needs no “tuning” and uses no potentially expensive “twists”, like combining by input dissimilarities.

5 Conclusions and Outlook

In this paper we examined several combination strategies for SVM. After combining SVM with a fixed input hypothesis and varying kernel functions it was found out, that a simple addition of output values leads to medium performance. But by including region information the classification accuracy rises dramatically. The price to pay is, that for some credit clients no classification is possible. Enhancing the number of incorporated models will reduce this number of “unclassified” clients and, especially in non standard situations of classification, we propose to introduce *default classes* to overcome this problem. Finally, adding several simple selections of outputs of *weak* base SVM classifiers (on variable subsets of data) yields misclassification accuracies comparable to extensively trained and selected single models. Training and validating a supervised combiner on the inputs of these simple combiners confirms and further improves this result.

References

1. Duin, R.P.W. and Tax, D.M.J. (2000): Experiments with Classifier Combining Rules. In: Kittler, J. and Roli, F. (Eds.): MCS 2000, LNCS 1857. Springer, Berlin Heidelberg, 16-19
2. Kuncheva, L.I. (2004): *Combining Pattern Classifiers: Methods and Algorithms*. Wiley 2004
3. Schebesch, K.B. and Stecking, R. (2005): Support Vector Machines for Credit Scoring: Extension to Non Standard Cases. In: Baier, D. and Wernecke, K.-D. (Eds.): *Innovations in Classification, Data Science and Information Systems*. Springer, Berlin, 498-505.
4. Schebesch, K.B. and Stecking, R. (2005): Support vector machines for credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56(9), 1082-1088.
5. Schebesch, K.B. and Stecking, R. (2006): Selecting SVM Kernels and Input Variable Subset in Credit Scoring Models, submitted to Proceedings of the 30th Annual Conference of the GfKI 2006.
6. Stecking, R. and Schebesch, K.B. (2006): Comparing and Selecting SVM-Kernels for Credit Scoring. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nürnberger, A., Gaul, W. (Eds.): *From Data and Information Analysis to Knowledge Engineering*. Springer, Berlin, 542-549

Nutzung von Data-Mining-Verfahren zur Indexprognose

Jonas Rommelspacher

Institut für Wirtschaftsinformatik, Philipps-Universität Marburg, Germany
rommelspacher@wiwi.uni-marburg.de

1 Problemstellung

Finanzmarktakeure müssen ihren Investitionsentscheidungen Erwartungen bzgl. der zukünftigen Marktentwicklung zugrunde legen. Sie stehen vor einer Entscheidung unter Unsicherheit und sind bestrebt mittels Prognoseverfahren die zukünftige Kursentwicklung möglichst gut vorherzusagen.

Die klassische Kapitalmarkttheorie basiert im Kern auf der Theorie der rationalen Erwartungen (Muth) und der Hypothese der effizienten Informationsverarbeitung von Finanzmärkten (Fama). In ihrer schwachen Form postuliert letztere, dass die gegenwärtigen Preise alle Informationen der Vergangenheit widerspiegeln und die zukünftigen Preise einem Random Walk folgen [3]. Weiter besagt diese Hypothese, dass es unmöglich ist, mit vergangenen Kursen zukünftige Kurse vorherzusagen [7]. Somit steht die Hypothese der effizienten Informationsverarbeitung im Gegensatz zu der o.g. Anforderung der Finanzmarktakeure. Es besteht allerdings die Vermutung, dass die klassischen linearen Modelle aufgrund von Nichtlinearitäten in Finanzmarktreihen versagen [11].

Neben der Tatsache, dass einige Korrelationen an Finanzmärkten kaum zu leugnen sind, gilt es mittlerweile als relativ gesichert, dass aufgrund von Nichtlinearitäten durchaus bis zu einem gewissen Grad erfolgreiche Prognosen erstellt werden können [4].

Letztendlich handelt es sich bei der Prognose zukünftiger Kurse, gleich ob dafür mathematische Modelle oder heuristischer Verfahren (Gleitende Durchschnitte, Technische Analyse) eingesetzt werden, um ein Mustererkennungsproblem [5]. Zur Erkennung von Datenmustern werden üblicherweise Data-Mining-Verfahren aus dem Forschungsgebiet der Künstlichen Intelligenz eingesetzt.

Ausgehend von der geschilderten Problematik soll in diesem Beitrag diskutiert werden, inwiefern Data-Mining-Verfahren für die Prognose von Aktienindizes eingesetzt werden können.

Der Beitrag ist folgendermaßen aufgebaut: In (2) wird das Framework der Untersuchung dargestellt, bevor in (3) mittels Künstlichen Neuronalen Netzen und Nearest-Neighbour-Verfahren eine Indexprognose des Deutschen Aktienindexes (DAX) durchgeführt wird. Da es sich bei den dargestellten Prognosen um Single-Chart-Analysen handelt, wird in (4) diskutiert, wie die Prognose aus (3) durch

Nutzung von Genetischen Algorithmen erweitert werden kann und wie etwa Entscheidungsbäume zur Entwicklung einer sinnvollen Anlagestrategie eingesetzt werden können. Der Beitrag schließt mit einem Fazit (5).

2 Lösungsansatz zur Indexprognose

Im Folgenden soll kurz das Framework der Untersuchung dargestellt werden. Empirisch gibt es Hinweise darauf, dass die Prognosegüte massiv davon abhängt, in welcher Phase (z.B. Bull- oder Bear-Markt) sich der zu prognostizierende Markt befindet [5]. Grundsätzlich liegt die Vermutung nahe, dass Finanzzeitreihen nicht-stationär und deshalb nicht sinnvoll durch ein Globalmodell zu beschreiben sind. Da es in dem Beitrag um die grundsätzliche Eignung von Data-Mining-Verfahren zur Prognose geht wurde zur Umgehung des Problems ein relativ kleiner Datensatz ausgewählt.

Die durch die Algorithmen zu lösende Lernaufgabe ist es, basierend auf den historischen Daten des Deutschen Aktienindexes vom 2004-08-06 bis 2006-07-20 (500 Handelstage), so genannte steigt/fällt-Prognosen durchzuführen.¹ Es soll also lediglich die Richtung des Kurses prognostiziert werden.

Zur Messung der Performancequalität soll die Trefferquote verwendet werden. Da Aktienkurse im Allgemeinen schwierig vorherzusagen sind werden in hohem Maße Fehlurteile bei der Prognose akzeptiert. Um die Problematik zu verdeutlichen sei betont, dass es bei steigt/fällt-Prognosen keinesfalls darum geht, eine möglichst 100%ige Genauigkeit zu erreichen, sondern möglichst weit oberhalb der Zufallsgrenze (Trefferquote von 50 %) zu liegen. Problematisch an der Kennzahl Trefferquote ist, dass große und kleine Gewinne/Verluste gleich gewichtet werden. Zur Umgehung dieses Problems gibt es in der Literatur verschiedene Vorschläge [10][5].

Neben der Performancequalität müssen die Ergebnisse auf Signifikanz getestet werden. Dies ist mittels eines χ^2 -Unabhängigkeitstests zu beurteilen. (Vgl. zur ausführlichen Beschreibung des Tests: [2]).

3 Indexprognose mit Data-Mining-Verfahren

3.1 Künstliche Neuronale Netze

Künstliche Neuronale Netze ahmen die Funktionsweise des menschlichen Gehirns nach und gelten als ein flexibles und robustes nichtlineares statistisches Verfahren. Die Aufgabe jedes Neurons besteht darin, die Informationen von seinem vorgelagerten Nachbarn zu empfangen, zu verarbeiten und das Ergebnis an seinen jeweils nächsten Nachbarn weiterzugeben. Die Menge der an einem neuronalen Netz beteiligten Neuronen lässt sich in disjunkte Teilmengen (Schichten) aufteilen. Zwischen Eingabeschicht und Ausgabeschicht liegen in der Regel verborgene Schichten von Zwischenneuronen. Sind Rückkopplungen zwischen den nicht Schichten zugelassen, spricht man von einem Feedforward-Netz, andernfalls handelt es sich um ein Feedback-Netz. Die Ausgestaltung des Künstlichen Neuronalen Netzes wird auch Topologie genannt.

¹ Datenquelle: <http://finance.yahoo.com>

Durch die Parallelverarbeitung und Lernfähigkeit bieten sich Neuronale Netze für zu lösende nicht-lineare Problemstellungen an [7]. So überrascht es auch wenig, dass bereits seit Ende der 80er Jahre Neuronale Netze zur Kursprognose von Aktien eingesetzt werden.

In einem Optimierungsprozess lernt das Neuronale Netz mittels Trainingsdaten und passt sich an die Inputfaktoren an. Im Anschluss daran muss die Vorhersagegüte des neuronalen Netzes an einer möglichst unabhängigen Testmenge getestet werden.

Für die zu lösende Aufgabe wurde der so genannte CCN- (Cascade Correlation Neural Net)-Algorithmus verwendet, der ein Netz je nach Aufgabenstellung mit keiner oder mehreren verborgenen Neuronenschichten inkrementell aufbaut. Die Performance der Testdaten im Vergleich zu der tatsächlichen Entwicklung ist in Tabelle 1 dargestellt.

Table 1. Prognoseergebnis Künstliches Neuronales Netz (Testdaten)

	true fällt	true steigt	Σ
pred fällt	16	47	63
pred steigt	9	78	87
Σ	25	125	150

Die Trefferquote beträgt 62,67 % was empirisch gesehen ein sehr guter Wert ist [10][1]. Zudem zeigt der χ^2 -Unabhängigkeitstest, dass das Ergebnis signifikant auf dem 1%-Niveau ist.

Das generierte Neuronale Netz erlaubt aufgrund der Black-Box-Eigenschaft kaum analytische Interpretationen zwischen Eingabe- und Ausgabewerten [11].

3.2 Neareast-Neighbour-Verfahren

Das Nearest-Neighbour-Verfahren versucht geometrische Segmente in einer Zeitreihe zu erkennen und prognostiziert die zukünftige Kursentwicklung dadurch, dass es versucht ein ähnliches geometrisches Segment zu erkennen wie das Segment, das genau vor der vorherzusagenden Periode liegt.

Der Ansatz der Kursprognose mittels Nearest-Neighbour ist der Technischen Aktienanalyse im Prinzip verwandt. Diese wird von Händlern bereits seit langem zur kurzfristigen Kursprognose verwendet [4].

Bei der Kursprognose des Deutschen Aktienindexes wurde das k-Nearest-Neighbour-Verfahren verwendet, d.h. die k ähnlichsten unter den gelernten Objekten werden zur Beantwortung der Frage *steigt oder fällt* herangezogen. Es wurde ebenso wie in (3.1) in Trainingsdaten und Testdaten unterteilt. Die Performance der Testdaten im Vergleich zu der tatsächlichen Entwicklung ist in Tabelle 2 dargestellt. Die Trefferquote liegt mit 50,67 % sehr nahe an der Zufallsgrenze und der χ^2 -Unabhängigkeitstest zeigt, dass das Ergebnis nicht signifikant ist. Zu ähnlichen Ergebnissen gelangen Rodriguez et al. [4] bei einer Nearest-Neighbour-Prognose des Nikkei-Indexes. Zemke [10] kommt dagegen zu dem Schluß, dass das Nearest-Neighbour-Verfahren gut zur Indexprognose geeignet ist.

Table 2. Prognoseergebnis Nearest-Neighbour-Verfahren (Testdaten)

	true fällt	true steigt	Σ
pred fällt	22	35	57
pred steigt	39	54	93
Σ	61	89	150

4 Weiterentwicklung der Indexprognose

Bislang wurde lediglich die Kursreihe, die auch vorhergesagt werden soll als Inputgröße verwendet. Da aber die Entwicklung von Finanzzeitreihen von vielen ökonomische Faktoren abhängt, erscheint eine Weiterentwicklung hin zu einer Multiple-Chart-Analyse sinnvoll.

Dabei ist unabhängig vom verwendeten Prognoseverfahren die Auswahl der Inputwerte entscheidend für die Prognosegüte. Neben Rankingsystemen bieten sich insbesondere Genetische Algorithmen für die Auswahl an [1]. Genetische Algorithmen verwenden eine Variante des Versuch-und-Irrtum-Prinzips und sind an die Evolutionstheorie von Darwin angelehnt, d.h. sie lösen das jeweilige Optimierungsproblem durch Selektion und Mutation. In der ersten Generation werden die zur Verfügung stehenden Inputvariablen mit Operatoren zufällig zu Regeln verknüpft. Jedes Regelset wird dann weiter auf seine Performance getestet. Die weniger fitten Variablen werden gelöscht, die besten werden neu kombiniert. Auch zufällige Veränderungen (Mutationen) sind vorgesehen. Daraufhin wird wieder selektiert bis ein Regelset entsteht, das den vorher definierten Anforderungen entspricht [6].

Neben Prognosen über die zukünftige Kursentwicklung benötigen Finanzmarktakteure Angaben dazu, wie sie mit den generierten Prognoseinformationen umzugehen haben. Auch hierfür existieren im Forschungsgebiet Data Mining Lösungsansätze: So bieten sich etwa Entscheidungsbäume für die Darstellung von aufeinander folgenden, hierarchischen Entscheidungsregeln an. Eine sehr einfache Entscheidungsregel wäre:

WENN der für $t+1$ vorhergesagte Aktienkurs größer (kleiner) als der Kurs in t ist, DANN kaufe (verkaufe) die Aktie. Die Position muss eine Minute später wieder geschlossen werden.

Exemplarisch wurde diese Regel in Abbildung 1 als binärer einstufiger Entscheidungsbaum dargestellt. Darüber hinaus ist es bei Multiple-Chart-Analyse durchaus denkbar, per Regelinduktion Handelsregeln zu generieren. Dabei wird aus einer Lernmenge ein effizienter Entscheidungsbaum generiert. Als Induktionsverfahren stehen etwa der ID3- oder der C4.5-Algorithmus zur Verfügung. Diese werden zur Generierung von Entscheidungsbäumen verwendet, wenn bei großer Datenmenge viele verschiedene Attribute von Bedeutung sind. Die Algorithmen sind zwar effizient, finden aber nicht zwingend den bestklassifizierenden Baum.

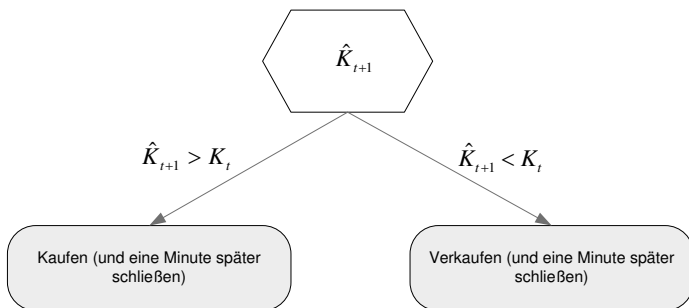


Fig. 1. Darstellung einer Entscheidungsstrategie

5 Fazit und weiterer Forschungsbedarf

In dem Beitrag wurde dargestellt, dass Data-Mining-Verfahren auf vielfältige Weise für die Finanzmarktanalyse und -prognose einsetzbar sind. Es wurde gezeigt, dass mit Künstlichen Neuronalen Netzen gute Prognoseergebnisse zu erzielen sind. Bei der Anwendung des Nearest-Neighbour-Verfahrens wurde jedoch deutlich, dass man keine übertriebenen Erwartungen an die Prognosequalität von nichtlinearen Verfahren haben darf. Offensichtlich weichen die Ergebnisse bzgl. der Eignung verschiedener Data-Mining-Verfahren zur Indexprognose auch je nach Design der Studie und dem entsprechenden Datensatz stark voneinander ab (Vgl. [dieser Beitrag] [10] [4]).

Es soll an dieser Stelle nicht verschwiegen werden, dass es in der Literatur (gerade von Vertretern der klassischen Finanzmarkttheorie) grundsätzliche Bedenken gegen eine zu datenorientierte Vorgehensweise gibt [8]. Eng damit verbunden ist die Gefahr des Overfittings, ein allgemeines Problem nichtlinearer Modelle. Dieses rührt daher, dass sich jede Funktion mathematisch durch die Wahl möglichst vieler Variablen beliebig genau nachbilden lässt. Allerdings läuft man damit Gefahr, dass der Prognosegehalt der Funktion gegen Null geht [8]. Dieses Problem ist hier zunächst zu vernachlässigen, da nur eine Inputvariable verwendet wurde.

Abschließend soll noch darauf eingegangen werden, welche weiteren, bisher nicht genannten Entwicklungsschritte folgen sollen: Zunächst muss die Frage geklärt werden, wie das oben bereits genannte Problem der Nichtstationarität gelöst werden kann. Zum Erkennen von verschiedenen Marktphasen werden in der Literatur bereits einige Ansätze diskutiert [9][5] und sollen zukünftig implementiert werden.

Zur Generierung besserer Prognoseergebnisse erscheint es zudem sinnvoll mehrere Data-Mining-Verfahren miteinander zu kombinieren, wie dies in (4) bereits angedeutet wurde. Ein integrierter Ansatz verspricht nach herrschender Meinung bessere Entwicklungsergebnisse [9].

References

1. Bohn A, Güting T, Mansmann T, Selle S (2003) MoneyBee: Aktienkursprognose mit künstlicher Intelligenz bei hoher Rechenleistung. *Wirtschaftsinformatik* 45: 325-333

2. Diebold F, Lopez J (1996) Forecast Evaluation and Combination. In: Maddala G, Rao, C (eds) Handbook of Statistics. North-Holland, Amsterdam: 241-268
3. Fama EF (1970) Efficient capital markets. A review of theory and empirical work. Journal of Finance 25: 383-417
4. Fernández-Rodríguez F, Sosvilla-Rivero S, Garca-Artiles MD (1999) Dancing with bulls and bears: Nearest-neighbour forecast for the Nikkei index. Japan and the World Economy 11 : 395-413
5. Heyder F, Zayer S (1999) Analyse von Kurszeitreihen mit Künstlichen Neuronalen Netzen und Competing Experts. Wirtschaftsinformatik 41:132-137
6. Kühn F (1998): Moderne Software-Werkzeuge. Optionsscheinmagazin: 91-93
7. Steiner M, Bruhns C (2002) Werpapiermanagement. 8. ed., Schäffer-Poeschel, Stuttgart
8. Schröder M (2002) Erstellung von Prognosemodellen. In: Schröder M (eds) Finanzmarkt-Ökonometrie: Basistechniken, Fortgeschrittene Verfahren, Prognosemodelle. Schäffer-Poeschel, Stuttgart: 397-465
9. Oh KJ, Kim K-J (2002) Analyzing stock market tick data using piecewise nonlinear model. Expert Systems with Applications 22: 249-255
10. Zemke S (1999) Nonlinear Index Prediction. Physica A 269: 177-183
11. Zschischang E (2005) Genetisches Programmieren als neues Instrumentarium zur Prognose makroökonomischer Größen. Doctoral Thesis, Christian-Albrechts-Universität, Kiel

Zur Entscheidungsunterstützung bei netzeffektbasierten Gütern

Karl-Heinz Lüke, Klaus Ambrosi, and Felix Hahne

Institut für Betriebswirtschaft Universität Hildesheim, Germany
www.uni-hildesheim.de/de/bwl.htm

Summary. Für marketingrelevante Fragestellungen und für die strategische Bewertung von Technologiealternativen ist die Untersuchung des Diffusionsverlaufs bei Netzeffektgütern, wie z.B. Telekommunikationsdiensten, von hohem Interesse. Es wird ein Entscheidungsunterstützungssystem vorgestellt, dass insbesondere Marketingvariable zur Ausgestaltung relevanter Merkmale von netzeffektbasierten Gütern unterstützt. Ausgehend von dynamischen Nutzenfunktionen wird ein Modellansatz auf Grundlage der Mastergleichung der Physik bzw. den Mittelwertgleichungen vorgestellt, der Wechselwahrscheinlichkeiten zwischen den Produktalternativen ableitet. Die Anwendungsrelevanz des Modellansatzes wird durch entscheidungsrelevante What-If-Analysen verdeutlicht.

1 Besonderheiten von Netzeffektgütern

Ein Ziel der Diffusionsforschung ist die Erklärung von zeitlichen Ausbreitungsverläufen von Innovationen bzw. Produkten.¹ Primäres Untersuchungsobjekt der klassischen Diffusionstheorie bilden Singulärgüter, die durch einen originären Nutzen gekennzeichnet sind. Davon zu unterscheiden sind im allgemeinen Netzeffektgüter, die sich in Netzeffektgüter i.e.S. und Systemgüter unterteilen lassen (vgl. Schoder (1995), S. 13). Netzeffektgüter weisen eine Reihe von Besonderheiten auf, die für die Diffusionsmodellierung berücksichtigt werden müssen (vgl. Schoder (1995), S. 147; vgl. Weiber (1992), S. 135). Neben dem originären Nutzen besitzen Netzeffektgüter i.e.S. noch einen derivativen Produktnutzen, der aus dem Verbreitungsgrad von komplementären Gütern am Markt bestimmt wird (z.B. CD-Player und CD). Kennzeichnendes Merkmal von Systemgütern ist der derivative Nutzen, der sich aus dem interaktiven Einsatz im Rahmen von Systemtechnologien ergibt (z.B. Telefon, E-Mail) (vgl. Schoder (1995), S. 11f.; vgl. Weiber (1992), S. 18).

¹ Siehe die umfassende Übersicht von Diffusionsmodellen bei Mahajan/Muller/Wind (2000).

2 Quantitative Beschreibung des Diffusionsmodells

2.1 Modellgleichung

Die Grundlage für das vorgestellte Diffusionsmodell bildet der Mastergleichungsansatz, der im Zusammenhang mit der Beschreibung physikalischer, chemischer und biologischer Systeme (vgl. Haken (1990)) Verwendung findet, wobei die Anzahl der ökonomischen und soziologischen Anwendungsgebiete der Mastergleichung zunimmt (vgl. Schoder (1995), S. 76). Die formale Herleitung der Mastergleichung ist z.B. bei Weidlich und Haag zu finden (vgl. Weidlich/Haag (1983), S. 58f.; vgl. Weidlich/Haag (1988), S. 321f.; vgl. Weidlich (2002), S. 311ff.).

Für die Mastergleichung werden sehr detaillierte Informationen benötigt, die in vielen Fällen aus empirischen Untersuchungen nicht zur Verfügung stehen. Daher werden die Mittelwertgleichungen herangezogen, die sich aus einer Vereinfachung der Mastergleichung bei unimodalen Wahrscheinlichkeitsverteilungen herleiten lassen (vgl. Weidlich/Haag (1988), S. 26f.; vgl. Weidlich/Haag (1983), S. 87f.; vgl. Weidlich (2002), S. 341ff.). Je nach Berücksichtigung von endogenen Einflussvariablen $m(t)$ (d.h. direkte Netzeffekte) kann entweder ein nichtlineares oder ein lineares Differentialgleichungssystem formuliert werden.

$$\frac{dn_k^\alpha(t)}{dt} = \underbrace{\sum_{\substack{i=1 \\ i \neq k}}^L p_{ik}^\alpha(m(t), s_i^\alpha(t), s_k^\alpha(t), a, b) \cdot n_i^\alpha(t)}_{\text{Zuflußterm}} - \underbrace{\sum_{\substack{j=1 \\ j \neq k}}^L p_{kj}^\alpha(m(t), s_k^\alpha(t), s_j^\alpha(t), a, b) \cdot n_k^\alpha(t)}_{\text{Abflußterm}} \quad (1)$$

- P : Anzahl der Konsumentengruppen: $\alpha = 1, \dots, P$
 L : Anzahl der Alternativen
 k, i, j : Indizes der Alternativen
 $s_i^\alpha(t)$: Vektor der Aktionsparameter der Alternative i der Konsumentengruppe α ,
 $s_i^\alpha(t) = (s_{i,1}^\alpha(t), \dots, s_{i,R}^\alpha(t))'$,
 R : Anzahl der Aktionsparameter, z.B. $s_{i,1}^\alpha(t)$: Preis;
 $s_{i,2}^\alpha(t)$: Service
 $m(t)$: Marktaufteilung bzw. Marktvektor
 $m(t) = (n_1^1(t), \dots, n_L^1(t), \dots, n_i^\alpha(t), \dots, n_1^P(t), \dots, n_L^P(t))'$

a, b :	Parameter der Transformationsfunktion φ
$n_k^\alpha(t), n_i^\alpha(t)$:	Anzahl der Adoptoren bzw. potentiellen Adoptoren der Alternative k, i zum Zeitpunkt t der Konsumentengruppe α
$n_k(t)$:	Anzahl der Adoptoren bzw. potentiellen Adoptoren der Alternative k zum Zeitpunkt t
	$n_k(t) = \sum_{\alpha=1}^P n_k^\alpha(t)$
$p_{ik}^\alpha(\dots)$:	Wechselwahrscheinlichkeit von Alternative i nach k der Konsumentengruppe α
$p_{kj}^\alpha(\dots)$:	Wechselwahrscheinlichkeit von Alternative k nach j der Konsumentengruppe α

2.2 Wechselwahrscheinlichkeiten

Der Nutzen $U_i^\alpha(m(t), s_i^\alpha(t))$ für eine Alternative setzt sich für netzeffektbasierte Güter aus einer Marktaufteilung $m(t)$ und den Aktionsparametern $s_i^\alpha(t)$ zusammen. Der Wertebereich der Nutzenfunktion ist auf den Bereich $0 \leq U_i^\alpha(m(t), s_i^\alpha(t)) \leq 1$ normiert. Der Wechsel von einer Alternative i zu einer Alternative j wird über die Nutzendifferenz $d_{ij}^\alpha(m(t), s_j^\alpha(t), s_i^\alpha(t))$ bestimmt (siehe (2)).

$$p_{ij}^\alpha(m(t), s_j^\alpha(t), s_i^\alpha(t), a, b) = \frac{\varphi(d_{ij}^\alpha(m(t), s_j^\alpha(t), s_i^\alpha(t)), a, b)}{\sum_{j=1}^L \varphi(d_{ij}^\alpha(m(t), s_j^\alpha(t), s_i^\alpha(t)), a, b)} \quad (2)$$

$p_{ij}^\alpha(\dots)$:	Wechselwahrscheinlichkeit
$p_{ii}^\alpha(\dots)$:	Wahrscheinlichkeit für Markentreue bei Alternative i
$U_i^\alpha(m(t), s_i^\alpha(t))$:	Nutzenfunktion der Alternative i , $0 \leq U_i^\alpha(m(t), s_i^\alpha(t)) \leq 1$
$d_{ij}^\alpha(\dots)$:	Nutzendifferenzen zwischen den Alternativen i und j $d_{ij}^\alpha(m(t), s_j^\alpha(t), s_i^\alpha(t)) =$ $\max\{0, U_j^\alpha(m(t), s_j^\alpha(t)) - U_i^\alpha(m(t), s_i^\alpha(t))\}, \forall i \neq j,$ $d_{ii}^\alpha(m(t), s_i^\alpha(t)) = U_i^\alpha(m(t), s_i^\alpha(t)), i = j$
$\varphi(\dots)$:	Transformationsfunktion mit den Parametern a, b

2.3 Transformationsfunktion

Eine typische Transformationsfunktion φ wird mit den Parametern a, b gewählt.

$$\varphi(d_{ij}^\alpha(m(t), s_i^\alpha(t), s_j^\alpha(t)), a, b) = 1 - e^{-a(d_{ij}^\alpha(m(t), s_i^\alpha(t), s_j^\alpha(t)))^b}, \quad (3)$$

$$\underline{a} \leq a \leq \overline{a}, \underline{b} \leq b \leq \overline{b}$$

$$\underline{a}, \underline{b}: \quad \text{Untere Grenzen, } \underline{a}, \underline{b} \geq 0$$

$$\overline{a}, \overline{b}: \quad \text{Obere Grenzen, } \overline{a}, \overline{b} \geq 0$$

3 Modellsimulation

Im Mittelpunkt der nachstehenden Ausführungen steht zum einen die Modellkalibrierung und zum anderen die Simulation entscheidungsrelevanter What-If-Analysen. Zielsetzung bei der Modellkalibrierung ist die Schätzung der Parameter a, b der Transformationsfunktion φ von (3) bei gegebenen Nutzen $U_i^\alpha(m(t), s_i^\alpha(t))$ über alle Perioden T , wobei a, b so zu wählen sind, dass die Summe der quadratischen Abweichungen über alle Zeitperioden ($t = 1, \dots, T$), für die Alternativen ($k, i, j = 1, \dots, L$) und für alle Konsumentengruppen ($\alpha = 1, \dots, P$) minimiert wird. Da das Datenmaterial meist nur in einer zeitdiskreten Form aus Marktuntersuchungen zur Verfügung steht, erfolgt eine Diskretisierung des Differentialgleichungssystems (vgl. Lücke (2006)). Für das vorliegende Optimierungsproblem² hat sich ein lokales Verbesserungsverfahren, speziell eine Pattern Search Methode, bewährt (vgl. Lewis/Torczon (1999)). Die Phase der What-If-Analyse, die auf der Modellkalibrierungsphase aufbaut, konzentriert sich auf die Untersuchung von Auswirkungen des Diffusionsverlaufs bei Änderung der Marketingvariable.

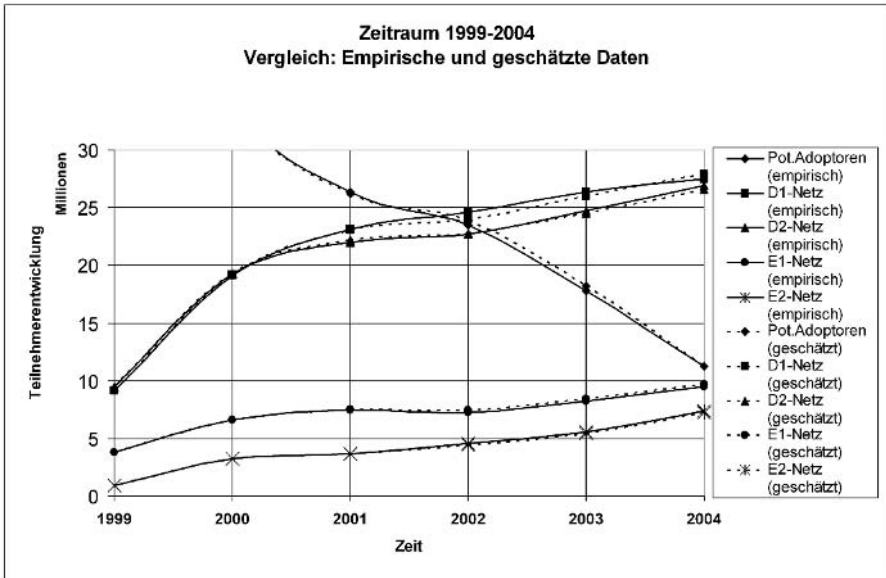


Fig. 1. Vergleich der empirischen und geschätzten Daten der Mobilfunkentwicklung (1999-2004) (vgl. Lücke (2006))

3.1 Modellkalibrierung

Das folgende Anwendungsbeispiel zeigt die Entwicklung des Mobilfunks in Deutschland. Die Daten stellen eine jahresbezogene, aggregierte Form der Teilnehmer-

² Siehe hierzu auch die bekannten Lösungsverfahren bei nichtlinearer Optimierung (vgl. Spellucci (1993); vgl. Großmann/Terno (1993); vgl. Geiger/Kanzow (1999)).

entwicklung in den Jahren 1992 bis 2004 vom D1-Netz, D2-Netz, E1-Netz und E2-Netz dar (vgl. Bundesnetzagentur (2005)). Als Optimierungsmethode wird die Pattern Search Methode auf Basis von Lewis und Torczon gewählt, wie sie im Programmpaket Matlab zur Verfügung steht. Das Ergebnis für einen zeitlichen Ausschnitt (1999-2004) der Anpassung des Modells an die empirischen Daten (Bestimmtheitsmaß $R_B^2 = 0,99$) zeigt Abbildung 1 (vgl. Lüke (2006)).

3.2 What-If-Analyse

Basierend auf der Kalibrierung des Modells mit der Pattern Search Methode an Zeitreihen werden verschiedene Strategien für die nächsten Zeitschritte untersucht. Dazu können im Rahmen von entscheidungsrelevanten What-If-Analysen die

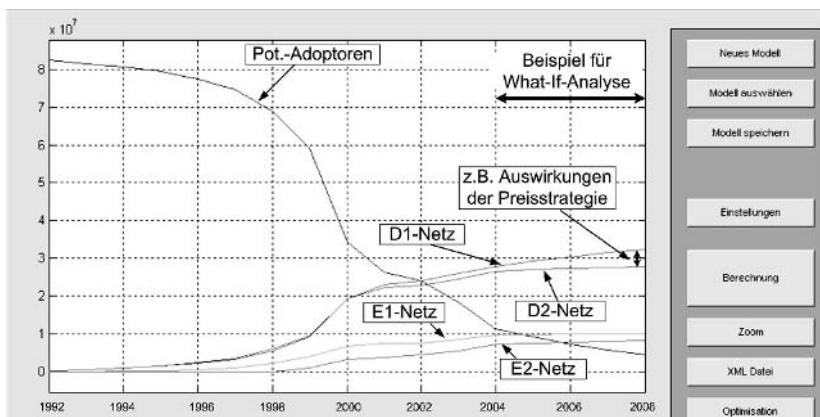


Fig. 2. Beispiel für What-If-Analyse (1): Preisstrategie A beim D1-Netz

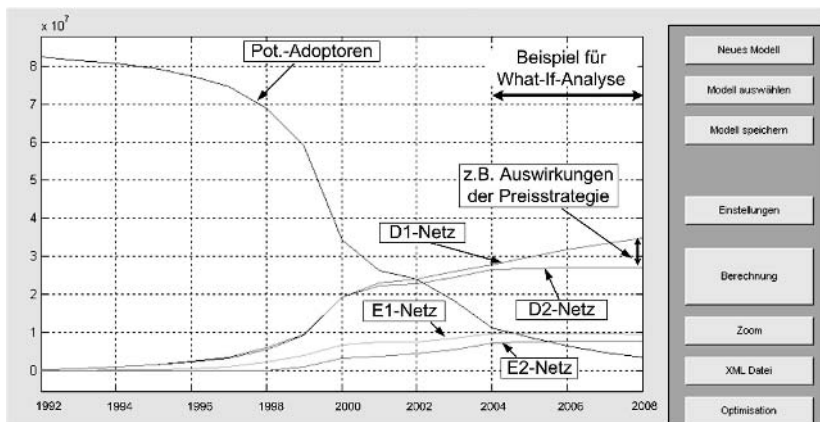


Fig. 3. Beispiel für What-If-Analyse (2): Preisstrategie B beim D1-Netz

Auswirkungen bei Veränderung einer oder mehrerer Aktionsparameter $s^\alpha(t)$ (z.B. Preisstrategie A bzw. B ab dem Jahr 2004) die Wirkungszusammenhänge in Abbildung 2 und 3 auf den Diffusionsverlauf veranschaulicht werden. Das Beispiel zeigt, daß Preisstrategie B (Abbildung 3) gegenüber der Preisstrategie A (Abbildung 2) beim D1-Netz zu einem höheren Teilnehmerwachstum führt.

References

1. Geiger, C. /Kanzow, C. (1999) Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben. Springer, Berlin [u.a.]
2. Großmann, C. /Terno, J. (1993) Numerik der Optimierung. Teubner, Stuttgart
3. Haken, H. (1990) Synergetik: Eine Einführung. Springer, 3. Auflage, Berlin
4. Lewis, R. M. /Torczon, V. (1999) Pattern Search Algorithms for Bound Constrained Minimization. In: SIAM Journal on Optimization, Bd. 9, Nr. 4, S. 1082-1099
5. Lücke, K.-H. (2006) Netzeffektbasierte Diffusion - Simulationsmodell zur Entscheidungsunterstützung für Analyse und Prognose. (Diss. Univ. Hildesheim, erscheint voraussichtlich im 4. Quartal 2006; Frankfurt am Main [u.a.]: Lang)
6. Mahajan, V./Muller, E. /Wind, Y. (2000) New-product diffusion models. Kluwer Acad. Publ., Boston [u.a.]
7. Bundesnetzagentur (2005) Webseite der Bundesnetzagentur.
URL: http://www.bundesnetzagentur.de/enid/5739e158c2b103444a227e33786c321c,0/Marktbeobachtung/Mobilfunkdienste_vw.html, Verifizierungsdatum: 15.11.2005.
8. Schoder, D. (1995) Erfolg und Mißerfolg telematischer Innovationen: Erklärung der "kritischen Masse" und weiterer Diffusionsphänomene. Dt. Univ.-Verl., Wiesbaden
9. Spellucci, P. (1993) Numerische Verfahren der nichtlinearen Optimierung. Birkhäuser, Basel [u.a.]
10. Weiber, R. (1992) Diffusion von Telekommunikation: Problem der kritischen Masse. Gabler, Wiesbaden
11. Weidlich, W. /Haag, G. (1983) Concepts and models of a quantitative sociology: The dynamics of interacting populations. Springer, Berlin [u.a.]
12. Weidlich, W. /Haag, G. (1988) Interregional migration: Dynamic theory and comparative analysis. Springer, Berlin [u.a.]
13. Weidlich, W. (2002) Sociodynamics: A systematic approach to mathematical modelling in the social sciences. Taylor & Francis, London

Discrete and Combinatorial Optimization

Nonserial Dynamic Programming and Tree Decomposition in Discrete Optimization

Oleg Shcherbina

Institut fuer Mathematik, University of Vienna, Austria
oleg.shcherbina@univie.ac.at

1 Introduction

Solving discrete optimization problems (DOP) can be a rather hard task. Many real DOPs contain a huge number of variables and/or constraints that make the models intractable for currently available solvers. There are few approaches for solving DOPs: tree search approaches (e.g., branch and bound), relaxation and decomposition methods. Large DOPs can be solved due to their special structure. Among decomposition approaches we can mention poorly known local decomposition algorithms using the special block matrix structure of constraints and half-forgotten nonserial dynamic programming algorithms which can exploit sparsity in the dependency graph of a DOP.

One of the promising approaches to cope with NP-hardness in solving DOPs is the construction of decomposition methods [7]. Decomposition techniques usually determine subproblems, whose solutions can be combined to create a solution of the initial DOP problem. Usually, DOPs from applications have a special structure, and the matrices of constraints for large-scale problems have a lot of zero elements (sparse matrices).

This paper reviews main results for local decomposition algorithms in discrete programming and establishes some links between them, tree decompositions and nonserial dynamic programming.

2 Nonserial Dynamic Programming

One of the promising ways to exploit sparsity in the dependency graph of an optimization problem is nonserial dynamic programming (NSDP) (BERTELE, BRIOSCHI [2], HOOKER [5]), which allows to compute a solution in stages such that each of them uses results from previous stages.

Classical dynamic programming has been proposed by R. Bellman as discrete analogue of the optimal control theory and is widely used in applications of operations research.

NSDP [2] appeared in 1960th but is poorly known to the optimization community.

This approach is used in Artificial Intelligence under the names "Variable Elimination" or "Bucket Elimination" [4]. Nonserial dynamic programming being a natural and general decomposition approach, considers a set of constraints and an objective function as recursively computable function. This allows to compute a solution in stages such that each of them uses results from previous stages.

An efficiency of this algorithms crucially depends on the interaction graph structure of a DOP. If the interaction graph is rather sparse or, in other words, has a relatively small induced width, then the complexity of the algorithm is reasonable.

We consider a DOP in the following form

$$F(x_1, x_2, \dots, x_n) = \sum_{k \in K} f_k(Y^k) \rightarrow \max \quad (1)$$

subject to the constraints

$$g_i(X^i) \leq 0, \quad i \in M = \{1, 2, \dots, m\}, \quad (2)$$

$$x_j \in D_j, \quad j \in \{1, \dots, n\}, \quad (3)$$

where

$$Y^k \subseteq \{x_1, x_2, \dots, x_n\}, X^i \subseteq \{x_1, x_2, \dots, x_n\}, R_i \in \{\leq, =, \geq\}. \quad (4)$$

Introduce an interaction graph [2] (dependency graph by HOOKER [5]) representing in natural way a structure of the DOP. The interaction graph will be denoted $G = (V, E)$ where a vertex $j \in V$ corresponds to a variable x_j of the DOP. Vertices 1 and 2 are adjacent (denoted $1 \sim 2$) if the variables x_1 and x_2 (corresponded to these vertices) appear together in the same member f_k of the objective function or in the same constraint i (in other words, if variables x_1 and x_2 are in the set Y^k or in the set X^i). Further, we shall use the notion of vertices that correspond one-to-one to variables. Let S_1, S_2 , and S be vertex sets of the graph, and j be a vertex of the graph. Introduce the following notions:

1. Adjacency of sets: $S_1 \sim S_2$ (read as " S_1 is adjacent to S_2 "), if there exists a vertex $j_1 \in S_1$ and a vertex $j_2 \in S_2$ such that $j_1 \sim j_2$.
2. $j \sim S$ means $\{j\} \sim S$.
3. Neighborhood of a set $S \subseteq V$, $Nb(S) = \bigcup_{v \in S} Nb(v) - S$.
4. Closed neighbourhood of a set $S \subseteq V$, $Nb[S] = Nb(S) \cup S$.

Consider a DOP and suppose without loss of generality that variables are eliminated in the order $x_1, \dots, x_n$. Using the NSDP scheme eliminate a first variable x_1 . This x_1 is in a set K_1 of objective function members:

$$K_1 = \{k \mid x_1 \in Y^k\}$$

and in a set of constraints with the indices U_1 :

$$U_1 = \{i \mid x_1 \in X^i\}.$$

Simultaneously with x_1 in objective function members Y^k , $k \in K_1$ and in constraints U_1 are variables from $Nb(x_1)$.

To the variable x_1 corresponds the following subproblem P_1 of DOP:

$$h_1(Nb(x_1)) = \max_{x_1} \left\{ \sum_{k \in K_1} f_k(Y^k) \mid g_i(X^i) \leq 0, i \in U_1, x_j \in D_j, x_j \in Nb(x_1) \right\}.$$

Then the initial DOP can be transformed in the following way:

$$\max_{x_1, \dots, x_n} \left\{ \sum f_k(Y^k) \mid g_i(X^i) \leq 0, i \in M, x_j \in D_j, j \in N \right\} =$$

$$\max_{x_1, \dots, x_{n-1}} \left\{ \sum_{k \in K-K_1} f_k(Y^k) + h_1(Nb(x_1)) \mid g_i(X^i) \leq 0, i \in M-U_1, x_j \in D_j, j \in N-\{1\} \right\}.$$

The last problem has $n-1$ variables; compared with the initial DOP from it were excluded constraints with the indices from U_1 and all objective function terms $\sum_{k \in K_1} f_k(Y^k)$ were excluded and there appeared a new objective function term $h_1(Nb(x_1))$. Due to this fact the interaction graph for the new problem changed: a vertex 1 was eliminated and its neighbours have become connected (due to the appearance a new term $h_1(Nb(x_1))$ in the objective). It can be noted that a graph generated by vertices from $Nb(x_1)$ is complete, i.e. is a clique. Denote a new interaction graph G^1 and find all neighbourhoods of variables in G^1 . NSDP eliminates the remaining variables one by one in an analogous manner. We have to store tables with optimal solutions at each stage of this process.

At the stage n of the described process we eliminate a variable x_n and find an optimal value of the objective function. Then a backward step of the NSDP procedure is performed using the tables with solutions.

The process of interaction graph transformation corresponding to the NSDP scheme is known as *Elimination Game* which was first introduced by PARTER [8] as a graph analog of Gaussian elimination.

3 Tree Decomposition Techniques

The search for graph structures appropriate for the application of dynamic programming caused a series of papers dedicated to tree decomposition research ([1], [3], [4], [6], [9]). Tree decomposition and the related notion of a treewidth (ROBERTSON, SEYMOUR [9]) play a very important role in algorithms, for many NP-complete problems on graphs that are otherwise intractable become polynomial time solvable when these graphs have a tree decomposition with restricted maximal size of cliques (or have a bounded treewidth [3]).

But only few papers on applications of this powerful tool in the area of DP exist.

Definition 1. Let $G = (V, E)$ be a graph. A tree decomposition of G is a pair $(T; X)$ with $T = (I; F)$ a tree and $X = \{X_i \mid i \in I\}$ a family of subsets of V , one for each node of T , such that

- (i) $\bigcup_{i=1 \in I} X_i = V$,
- (ii) for every edge $(u, v) \in E$ there is an $i \in I$ with $u \in X_i, v \in X_i$,
- (iii) for all $i, j, l \in I$, if $i < j < l$, then $X_i \cap X_l \subseteq X_j$.

4 Dynamic Programming with Tree Decompositions and Local Decomposition Algorithms for Discrete Programming

To describe how tree decompositions are used to solve problems with dynamic programming, let us assume we find a tree decomposition of a graph G . Since this tree decomposition is represented as a rooted tree T , the ancestor/descendant relation is well-defined. We can associate to each supernode X the subgraph of G made up by the vertices in X and all its descendant supernodes, and all the edges between those vertices.

Starting at the leaves of the tree T , we can compute information typically stored in a table, in a bottom-up manner for each bag until we reach the root. This information is sufficient to solve the subproblem for the corresponding subgraph. To compute the table for a node of the tree decomposition, we only need the information stored in the tables of the children (i.e. direct descendants) of this node. This is typical for dynamic programming. The problem for the entire graph can then be solved with the information stored in the table of the root of T . This approach is explained much more elaborately in [3]. Often the size of the tables is exponential in the width of the tree-decomposition used, which in turn is crucial for the running times computing for these tables. We therefore would like to use tree decompositions of width as small as possible. The smallest possible width is the treewidth of the graph and is NP-hard to compute in general.

During the study of complex objects it is not always possible (and expedient) to obtain (or to calculate) complete information about the object as a whole; therefore it is of interest to obtain information about the object, to examine it in parts, i.e., locally.

Local decomposition algorithms (LA) [12], [10] for the solution of DOPs use a concept of a neighborhood of the variables of the DOP and finding an optimal solution of the DOP is achieved with the aid of the study of neighborhoods and the decomposition of the initial DOP on this basis.

Consider the DOP Z with binary variables. Each step of the LA consists of the block elimination, and the change of neighborhoods of variables. Then the LA at the r th step is solving the optimization problem whose set of variables is determined by the neighborhood S_r , the set of constraints by the neighborhood U_r .

Consider a LA for solution of DOP problems with a tree-like structure, i.e., problems in which it is possible to find the set of the neighborhoods of different variables so that one variable can belong to two neighborhoods only and the graph of intersections of these neighbourhoods is a tree. It is clear that such a structure is a tree decomposition and can be obtained with the aid of the known tree decomposition algorithms ([3], [6]). The LA solves this DOP, moving bottom-up, i.e., from the neighborhoods corresponding to leaves of the tree, to the neighborhood corresponding to the root of the tree.

An estimate of the efficiency of the LA when solving a DOP with a tree-like structure which is characterized by the tree D with weights of the vertices n_r and weights of the edges $n_{r,r'}$, with binary variables and k blocks has the form

$$E(n, k, D) = \sum_r 2^{N_r} \phi(n_r), \quad (5)$$

where $\phi(n_r)$ is the estimate of the efficiency of the DP algorithm which solves DP problems corresponding to the blocks B_r ; here

$$N_r = n_{p_r r} + \sum_{r' \in J_r} n_{r r'}. \quad (6)$$

When using a complete enumeration algorithm to solve the DOP corresponding to the blocks, $\phi(n_r) = 2^{n_r}$. Using this estimate it is possible to show that a transfer from a tree-like structure to a quasiblock one has obtained by joining blocks is not expedient.

It is interesting to obtain the extremal and average values of the estimate of efficiency of the LA for the specified structure D and parameters n, k . Proofs of these propositions are in [11].

Proposition 1. *Writing $d_{r^*} = \max d_r$, $d_{r_1(r^*)} = \max_{r \in J_{r^*} \cup p_{r^*}} d_r$ we have*

$$\max_{n_r, n_{r r'}} E(n, k, D) = 2^{n-2k+2} (2^{d_{r^*}} + 2^{d_{r_1(r^*)}}) + \sum_{r \neq r^*, r_1(r^*)} 2^{d_r+1}.$$

Proposition 2. *If $n > k \max d_r - k + 1$, then*

$$\min E \geq k 2^{(n+k-1)/k}.$$

Proposition 3. *If $\phi(n_r) = 2^{n_r}$, then*

$$\langle E(n, k) \rangle \sim C_1(k) 2^n / n^{2k-2-\max d_r}, \quad n \rightarrow \infty, k > 2.$$

Proposition 4. *(necessary condition for the existence of a tree-like structure)*

If H is an $m \times n$ matrix with N_0 zero elements, then for it to have a tree-like structure with k blocks it is necessary that $n \geq 2k - 1, m \geq k$ and

$$N_0 \geq (k-2)(2m+n-2k+d^*+2) - m(d^*-2) - 3k + 2d^* + 4, k > 2.$$

5 Conclusion and Future Work

In this paper the class of discrete optimization problems with a tree-like structure is considered and based on nonserial dynamic programming (NSDP), and algorithms are proposed for solving this class of problems. Due to the fact that an NSDP algorithm is exact it has the known drawback of worst case exponential time performance. To cope with this difficulty, we plan to consider in the future questions of application of a postoptimality analysis in the NSDP (each step of the proposed algorithm consists of solving a series related integer programming problems), design of approximate versions of NSDP algorithms, and the use of relaxations and fathoming in the algorithmic scheme of the NSDP.

References

1. Arnborg S, Lagergren J, Seese D (1991) Easy problems for tree-decomposable graphs, J. of Alg. 12:308–340

2. Bertele U, Brioschi F (1972) Nonserial Dynamic Programming, Academic Press. New York
3. Bodlaender HL (1997) Treewidth: Algorithmic techniques and results. In: Privara, L. et al.(ed) , Mathematical foundations of computer science 1997. Proceedings. Berlin: Springer. Lect. Notes Comput. Sci. 1295:19–36
4. Dechter R, El Fattah Y (2001) Topological parameters for time-space tradeoff. Artificial Intelligence 125: 93–118
5. Hooker JN (2000) Logic-based methods for optimization: combining optimization and constraint satisfaction. John Wiley & Sons
6. Kloks T (1994) Treewidth: Computations and Approximations, Lecture Notes in Computer Science 842, Springer-Verlag
7. Nowak I (2005) Lagrangian decomposition of block-separable mixed-integer all-quadratic programs. Math. Programming, 102: 295–312.
8. Parter S (1961) The use of linear graphs in Gauss elimination. SIAM Review, 3:119–130
9. Robertson N, Seymour PD (1986) Graph minors. II. Algorithmic aspects of tree width. J.Algorithms, 7:309–322
10. Shcherbina OA (1983) On local algorithms of solving discrete optimization problems. Problems of Cybernetics. Moscow, 40:171–200
11. Shcherbina OA (1985) Local algorithms for block-tree problems of discrete programming. U.S.S.R. Comput. Math. Math. Phys. 25:114–121
12. Zhuravlev YuI, Finkelshtein YuYu (1965) Local algorithm for integer linear programming problems, Problems of Cybernetics 14, Moscow (in Russian).

Mixed-Model Assembly Line Sequencing Using Real Options

Alireza Rahimi-Vahed, Masoud Rabbani, Reza Tavakkoli-Moghaddam,
Fariborz Jolai, and Neda Manavizadeh

Department of Industrial Engineering, University of Tehran, Iran
payam.engmail@ut.ac.ir

1 Introduction

Monden [11] defined two goals for the mixed-model assembly line sequencing problem: (1) Leveling the load on each station on the line, and (2) Keeping a constant rate of usage of every part used by the line. To handle these problems, Goal chasing I and II (GC- I and GC- II) were developed by Toyota corporation. Miltenburg [9] developed a nonlinear programming for the second goal and solved the problem by applying two heuristic procedures. Miltenberg et al [10] solved the same problem with a dynamic programming algorithm. The objective considered by Bard et al [1] was the minimization of overall line length. Bard et al [2] used Tabu search (TS) algorithm to solve a model involving two objectives: minimizing the overall line length and keeping a constant rate of part usage. Hyun et al [4] addressed three objectives: minimizing total utility work, keeping a constant rate of part usage and minimizing total setup cost. This problem was solved by proposing a new genetic evaluation. McMullen [6] considered two objectives: minimizing number of setups and keeping a constant rate of part usage. He solved this problem with a TS approach. McMullen [7,8] has also solved the same problem by using genetic algorithm, and ant colony optimization, respectively.

The structure of this paper is as follows: In section 2, we present a detailed description of the mixed-model assembly line. In section 3, we show how real options can be applied to product-mix flexibility. In section 4, in order to determine the desired sequences, we consider three objectives simultaneously as follows: 1) total utility work cost, 2) total production rate variation cost, and 3) total setup cost. In Section 5, we propose the genetic algorithm and memetic algorithm. In Section 6, experimental results are given and various test problems are provided. Finally, we present our conclusions in Section 7.

2 Mixed-Model Assembly Line (MMAL)

An MMAL considered in this paper is a conveyor system moving at a constant speed (v_c). The line is partitioned into J stations. It is assumed that the stations are all closed types. The worker moves downstream on the conveyor while performing

his/her tasks to assemble a product. On completion of the job, the worker moves upstream to the next product. The design of the MMAL involves several issues such as determining operator schedules, product mix, and launch intervals. First, early start schedule is more common in practice and is used in this paper (Hyun, et. al. 1998). Second, the minimum part set (MPS) production, which this strategy is widely accepted in mixed model assembly lines, is also used in this paper. MPS is a vector representing a product mix, such that $(Q_1, Q_2, \dots, Q_M)$; where M is the total number of models. This strategy operates in a cyclical manner. The number of products produced in one cycle is given by $(I = \sum_{i=1}^M Q_i)$. The capacity of this line is limited with the maximum value, (C_{max}) , so that $(C_{max} \leq I)$. Third, the launch interval (γ) is set to $(\frac{T}{I \times J})$, in which T is the total operation time required to produce one cycle of MPS products.

3 Real Options and Product-Mix Flexibility

3.1 A Real Options Model

In this section, we will provide a real options model to evaluate product-mix flexibility. Investing in this line will give us numerous European options to produce so as to maximize the contribution margin.

$$V_k(T_K) = \max \sum_{i=1}^M \max[(P_i - VC_i) \cdot Q_{ik} + (P_i - OC_i) \cdot OQ_{ik} - SC_i, 0] \quad (1)$$

$$S.T. \quad Q_{ik} + OQ_{ik} = D_{ik} \quad i = 1, \dots, N \quad k = 1, \dots, K \quad (2)$$

$$\sum_{i=1}^M Q_{ik} \leq C_{max} \quad k = 1, \dots, K \quad (3)$$

$$Q_{ik}, OQ_{ik} \geq 0 \quad \text{and} \quad \text{Integer} \quad (4)$$

where $V_k(T_K)$ is the total contribution margin of the production decision at pre-set points in time (T_k) , $k \in [1, K]$ and $T_1 < T_2 < \dots < T_K$, P_i is the price of product i , VC_i is the variable cost of product i on mixed-model assembly line, OC_i is outsourcing cost of product i , OQ_{ik} is outsourced quantity of product i at time (T_k) , SC_i is the sequence-independent set-up cost for product i on mixed-model assembly line and C_{max} is the maximum capacity of line.

This real option model depends on value of several underlying assets. To evaluate these non-traded and non-financial underlying assets, it can be shown that The Q-Dynamics of the process D_i is as follows:

$$dD_i = (r - \delta)D_i dt + D_i \sum_{j=1}^M \sigma_{ij} dW_j; \quad (dW_i dW_j = 0 \quad i \neq j) \quad (5)$$

In Eq.(5), r is the risk-free rate of return, σ the parameter dependent on shortfall in rate of return, ρ_{ij} the correlation between demand of product i and demand of j , σ_i the volatility rate of demand of product i , and dW_j denotes the increment of a Wiener process where the mean is equal to zero and the variance is equal to the time increment dt .

To estimate the value of the option at time T_k , the following steps are implemented:

- 1) Simulate "risk-free" demand, under equation (2), with Monte Carlo simulation approach for all products giving a demand vector D .
- 2) For a given D , optimize production so as to maximize the value of the production schedule.
- 3) Perform the steps above n number of times.
- 4) Estimate the expected value of the option, i.e. the mean of the pay-off over all simulation runs.
- 5) Discount the expected value with the risk-free rate to get the present value of the option. This procedure is repeated for all available options and an aggregate value of all options can thereafter be determined.

4 Sequencing Objective Functions

4.1 Minimizing Total Utility Work Cost

The utility work is typically handled by the use of utility workers who assist the regular workers during work overload.

4.2 Minimizing Total Production Rate Variation Cost

One basic requirement of JIT systems is continual and stable part supply. Since this can be realized when the demand rate of parts is constant over time, the objective is important to a successful operation of the system. Therefore, the objective can be achieved by matching demand with actual production.

4.3 Minimizing Total Setup Cost

In many industries, sequence-dependent setups are considered as an important item in assembly operations.

5 MMAL Solution Algorithms

5.1 The Proposed Genetic Algorithm

5.1.1 Initialization

In the first step, a number of feasible solutions have to be introduced as the initial population (generation 0). To do so, a set of N chromosomes are randomly generated.

5.1.2 Parent Selection

A number of individuals are selected according to their fitness to form next generation. This is carried out using roulette wheel selection without replacement.

5.1.3 Crossover

In this paper, the order crossover was selected for the model.

5.1.4 Inversion and Mutation

In this study, to prevent the slow convergence of algorithm, inversion and swapping mutation was also used.

5.2 The Proposed Memetic Algorithm

The GA part in the proposed memetic algorithm is similar to the proposed genetic algorithm that was explained in section 5.1. Hence, only the difference, in a hybrid local search are discussed in this section.

If a new chromosome is produced in the initial population, then it is improved using the pairwise exchange procedure (XP). If a chromosome is provided by the mutation, then the insert procedure (IP) is utilized. These local searches must be performed until the fitness function value is unchanged for η consecutive times.

6 Experimental Results

The experiments are implemented in two folds: first, for small-sized problems, the other for large-sized ones. For both of these experiments, we consider the following assumptions: **-Real option's assumptions:** (1) All demands follow geometric Brownian motion, (2) The analysis will be carried out using a correlation between products equal to zero and a relative standard deviation of 0.25 , (3) Risk-free rate of return and the rate of return difference, δ , will be considered 0.05 and zero, respectively, and (4) Production strategy will be considered on a daily basis and for each day, we will carry out 1000 simulations. **-Sequencing assumptions:** (1) All relative weights have the same importance and are equal to 1, (2) The velocity of conveyor, (v_c) , is set to 1 for the sake of computational convenience, and (6) Each experiment is repeated 10 times. **-Genetic and Memetic algorithm's assumptions:** (1) The number of new offspring is fixed to 0.50 of the population size, (2) The ratio of ox-crossover, inversion and mutation is set to 1, 0.5, 0.076, respectively, and (3) The ratio of elitism is set to 0.5.

Table 1. Production Characteristics

Product	Average Daily Demand	Price	Variable Cost	Outsourcing Cost	Set-Up Cost
1	4	10	6	14	3
2	4	10	6	15	6
3	4	7	6	6	7
4	4	6	3	5	8
5	3	8	4	7	8
6	3	9	4	14	5
7	3	7	4	19	6
8	3	10	4	14	5

6.1 Small-Sized Problems

The problem considered is with four workstations and eight product models. The major production Characteristics of these products are presented by Table 1.

To establish the desired product-mix, real options approach has been carried out for 10 production days; ($k = 1,..,10$) . For this purpose, the maximum capacity of line, C_{max} , is set to 10 number of products. Table 2 shows the obtained results as follows: To determine the best sequence for each day, the proposed genetic algorithm

Table 2. Real options outputs

Day	Best Option	Desired Product-Mix	Outsourced quantity
1	(4,3,6,6,4,3,4,3)	(0,0,0,0,0,3,4,3)	(4,3,6,6,4,0,0,0)
2	(3,6,6,4,3,5,5,4)	(1,0,0,0,0,0,5,4)	(2,6,6,4,3,5,0,0)
3	(5,5,5,6,4,3,4,4)	(1,1,0,0,0,0,4,4)	(4,4,5,6,4,3,0,0)
4	(5,6,6,4,4,4,3,3)	(3,0,0,0,0,4,0,3)	(2,6,6,4,4,0,3,0)
5	(5,4,5,6,4,3,4,4)	(1,1,0,0,0,0,4,4)	(4,3,5,6,4,3,0,0)
6	(4,4,6,5,4,3,5,5)	(0,0,0,0,0,0,5,5)	(4,4,6,5,4,3,0,0)
7	(5,5,5,4,4,4,3,2)	(1,0,0,0,0,4,3,2)	(4,5,5,4,4,0,0,0)
8	(5,3,6,5,4,3,4,4)	(1,1,0,0,0,0,4,4)	(4,2,6,5,4,3,0,0)
9	(4,5,7,3,3,4,4,4)	(1,1,0,0,0,0,4,4)	(3,4,7,3,3,4,0,0)
10	(4,4,5,5,4,3,4,5)	(1,0,0,0,0,0,4,5)	(3,4,5,5,4,3,0,0)

and the proposed memetic algorithm are applied to the above problems. For each instance, the results obtained from these two algorithms are compared with the Lingo 6 software. Figure 1 presents obtained results, Schematically.

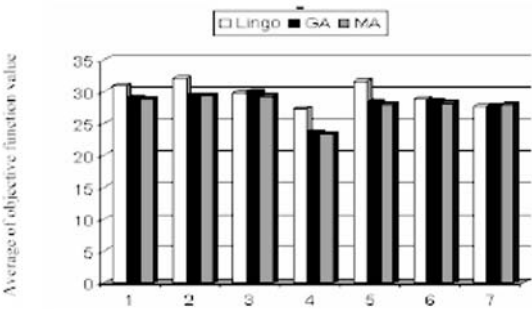


Fig. 1. Comparison of average of objective function value

6.2 Large-Sized Problems

In large-sized problems, The number of stations is fixed to 10 and mixed-model assembly line can produce eight product models. All major production characteristics of products are similar to Table 1, except the average daily demand that is

changed to (3,4,4,4,5,5,3). Table 3 shows the results obtained by real options approach. Figure 2 presents the comparison of average of the objective function value obtained by genetic algorithm and memetic algorithm, respectively.

Table 3. Real options outputs

Day	Best Option	Desired Product-Mix	Outsourced quantity
1	(3,5,7,4,4,6,7,4)	(3,0,0,0,0,6,7,4)	(0,5,7,4,4,0,0,0)
2	(4,6,5,4,6,7,4,2)	(1,6,0,0,0,7,4,2)	(3,0,5,4,6,0,0,0)
3	(2,6,6,4,5,7,8,4)	(1,0,0,0,0,7,8,4)	(1,6,6,4,5,0,0,0)
4	(3,4,7,5,3,5,8,5)	(2,0,0,0,0,5,8,5)	(1,4,7,5,3,0,0,0)
5	(5,5,4,4,8,6,4,2)	(3,5,0,0,0,6,4,2)	(2,0,4,4,8,0,0,0)
6	(3,4,5,5,6,4,7,4)	(1,4,0,0,0,4,7,4)	(2,0,5,5,6,0,0,0)
7	(4,7,6,3,6,6,5,3)	(0,6,0,0,0,6,5,3)	(4,1,6,3,6,0,0,0)
8	(4,4,5,6,7,5,6,3)	(2,4,0,0,0,5,6,3)	(2,0,5,6,7,0,0,0)
9	(3,4,5,5,7,4,6,4)	(2,4,0,0,0,4,6,4)	(1,0,5,5,7,0,0,0)
10	(4,7,6,4,5,8,6,4)	(2,0,0,0,0,8,6,4)	(2,7,6,4,5,0,0,0)

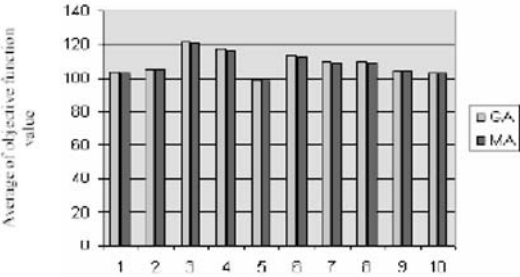


Fig. 2. Comparison of the average of the objective function value

7 Conclusions

In this paper, we used a real options model to evaluate productmix flexibility to recognize the products'quantity that must be satisfied by mixed-model assembly line. Then, in order to determine a desired sequence, We have considered three objectives simultaneously; minimizing the total utility work cost, total production rate variation cost, and total setup cost.Since the problem was NP-hard, genetic algorithm (GA) and memetic algorithm (MA) were proposed to solve it. Extensive computational experiments were performed for the proposed algorithms. The results have shown that the proposed memetic algorithm performs better than the proposed genetic algorithm.

References

1. Bard, J. F., Dar-El, E. M., and Shtub, A. (1992). An analytic framework for sequencing mixed model International Journal of Production Research, 30, 35–48.
2. Bard, J. F., Shtub, A., Joshi, S. B. (1994). Sequencing mixed-model assembly lines to level parts usage and minimize the length. International Journal of Production Research, 32, 2431–2454.
3. Bengtsson, J., Olhager, J., (2001). Valuation product-mix flexibility using real options, International Journal of Production Economics, 78, 13–28.
4. Hyun, C.J., Kim, Y., Kim, Y.K., (1998). A genetic algorithm for multiple objective sequencing problems in mixed model assembly lines. Computers and Operations Research. 25 (7/8), 675–690.
5. Kulatilaka N., (1995). The value of flexibility: A general model of real options, in: L. Trigeorgis (Ed.), Real Options in Capital Investment: Models, Strategies, and Application, Praeger, New York.
6. McMullen, P.R., (1998). JIT sequencing for mixed-model assembly lines with setups using tabu search. Production Planning and Control 9, (5), 504–510.
7. McMullen, P.R., (2001a). An efficient frontier approach to addressing JIT sequencing problems with setups via search heuristics. Computers and Industrial Engineering . 41, 335–353.
8. McMullen, P.R., (2001c). An ant colony optimization approach to addressing a JIT sequencing problem with multiple objectives. Artificial Intelligence in Engineering. 15, 309–317.
9. Miltenburg, J., (1989). Level schedules for mixed-model assembly lines in just-in-time production systems. Management Science. 35 (2), 192–207.
10. Miltenburg, J., Steiner, G., Yeomans, S., (1990). A dynamic programming algorithm for scheduling mixed-model just-in-time production systems. Mathematical Computation Modeling . 13, 57–66.
11. Monden, Y., (1983), Toyota production system . Institute of industrial engineers Press, Atlanta, GA .

A New Approach for Mixed-Model Assembly Line Sequencing

Masoud Rabbani, Alireza Rahimi-Vahed, Babak Javadi, and
Reza Tavakkoli-Moghaddam

Department of Industrial Engineering, University of Tehran, Iran
`payam.rahimivahed@gmail.ut.ac.ir`

Summary. This paper presents a fuzzy goal programming approach for solving a multi-objective mixed-model assembly line sequencing problem in a just-in-time production system. A mixed-model assembly line is a type of production line that is capable of diversified small lot production and is able to respond promptly to sudden demand changes for a variety of models. Determining the sequence of introducing models to such an assembly line is of particular importance for the efficient implementation of just-in-time (JIT) systems. In this paper, we consider three objectives, simultaneously: minimizing total utility work, total production rate variation, and total setup cost. Because of existence conflicting objectives, we propose a fuzzy goal programming based approach to solve the model. This approach is constructed based on the desirability of decision maker (DM) and tolerances considered on goal values. To illustrate the behavior of the proposed model, some of instances are solved optimally and computational results reported.

1 Introduction

A mixed-model assembly line is a type of production line that is capable of diversified small lot production. The effective utilization of mixed-model assembly line requires to solve two problems in sequential manner: (1) Line design and balancing, and (2) Determination of the production sequence for different models. Monden [10] defines two goals for the sequencing problems: (1) Leveling the load on each station on the line, and (2) Keeping a constant rate of usage of every part used by the line. To handle these problems, Goal chasing I and II (GC- I and GC- II) are developed by Toyota corporation. Miltenburg [8] developed a nonlinear programming for the second goal and solved the problem by applying two heuristic procedures. Miltenburg et al [9] solved the same problem with a dynamic programming algorithm. The objective considered by Bard et al [1] was the minimization of overall line length. Bard et al [2] used Tabu search (TS) algorithm to solve a model involving two objectives: minimizing the overall line length and keeping a constant rate of part usage. Hyun et al [3] addressed three objectives: minimizing total utility work, keeping a constant rate of part usage and minimizing total setup cost. This problem

was solved by proposing a new genetic evaluation. McMullen [4] considered two objectives: minimizing number of setups and keeping a constant rate of part usage. He solved this problem with a TS approach. McMullen [5,6] has also solved the same problem by using genetic algorithm, and ant colony optimization, respectively.

In this paper, we consider three objectives simultaneously as follows: 1) total utility work, 2) total production rate variation, and 3) total setup cost. The main purpose of this paper is to apply the fuzzy goal programming methodology where simultaneous minimization of the above mentioned objectives is desired. The structure of this paper is as follows. In Section 2, we present a detailed description of the mixed-model assembly line. In Section 3, a fuzzy goal programming based algorithm is proposed. In Section 4, experimental results are given and various test problems are provided. Finally, we present our conclusions in Section 5.

2 Multi-Objective Sequencing Problem in MMAL

2.1 Mixed-Model Assembly Line

An MMAL considered in this paper is a conveyor system moving at a constant speed (v_c). The line is partitioned into J stations. It is assumed that the stations are all closed types. The worker moves downstream on the conveyor while performing his/her tasks to assemble a product. On completion of the job, the worker moves upstream to the next product.

The design of the MMAL involves several issues such as determining operator schedules, product mix, and launch intervals. First, early start schedule is more common in practice and is used in this paper ([4]). Second, the minimum part set (MPS) production, which this strategy is widely accepted in mixed model assembly lines, is also used in this paper. MPS is a vector representing a product mix, such that $(d_1, \dots, d_m) = (D_1/h, \dots, D_m/h)$; where M is the total number of models, D_M is the number of products of model type m which needs to be assembled during the entire planning horizon and h is the greatest common divisor or highest common factor of $D_1, D_2, \dots, D_M$. This strategy operates in a cyclical manner. The number of products produced in one cycle is given by $I = \sum_{i=1}^M d_i$. Obviously, h times the repetition of producing the MPS products can meet the total demand in the planning horizon. Third, the launch interval (γ) is set to $(\frac{T}{I \times J})$, in which T is the total operation time required to produce one cycle of MPS products ([4]).

2.2 Sequencing Objective Functions

Minimizing Total Utility Work

The utility work is typically handled by the use of utility workers who assist the regular workers during work overload.

Minimizing Total Production Rate Variation

One basic requirement of JIT systems is continual and stable part supply. Since this can be realized when the demand rate of parts is constant over time, the objective is important to a successful operation of the system. Therefore, the objective can be achieved by matching demand with actual production.

Minimizing Total Setup Cost

In many industries, sequence-dependent setups are considered as an important item in assembly operations.

3 Fuzzy Goal Programming Approach

3.1 Background

When vague information related to the objectives is presented, the problem can be formulated as a fuzzy goal-programming problem. A typical fuzzy mixed-integer goal programming problem (f-MIGP) formulation can be stated as follows:

Find x_i

To satisfy

$$Z_l(x_i) = \tilde{Z}_l \quad (l = 1, \dots, L) \quad (1)$$

$$h_j(x_i) \leq d_j \quad (j = 1, \dots, J) \quad (2)$$

$$s_k(x_i) = c_k \quad (k = 1, \dots, K) \quad (3)$$

$$x_i \geq 0 \quad (i = 1, \dots, n) \quad (4)$$

where, $Z_l(x_i)$ is the l^{th} goal constraint. $h_j(x_i)$ is the j^{th} inequality constraint. $s_k(x_i)$ is the k^{th} equality constraint. $\tilde{Z}_l$ is the target value of the l^{th} constraint. d_j is the available resource of the j^{th} inequality constraint. c_k is the available resource of the k^{th} equality constraint.

In relations given in Eq. (1), the symbol " $\simeq$ " indicates the fuzziness of the goal. It represents the linguistic term "about" and it means that $Z_l(x_i)$ should be in the vicinity of the aspiration. The l^{th} fuzzy goal signifies that the decision maker will be satisfied even for values slightly greater than (or lesser than) up to a stated deviations signified by the tolerance limit. The j^{th} system constraint $h_j(x_i) \leq d_j$ and the k^{th} system constraint $s_k(x_i) = c_k$ are assumed to be crisp.

Let p_l be the maximum tolerance limit to g_l determined by the decision maker (DM). Thus, using the concept of fuzzy sets [11], the membership function of the objective functions can be defined as follows:

$$if Z_l(X) < g_l \rightarrow \mu_{Z_l}(x) = 1 \quad (5)$$

$$if g_l \leq Z_l(X) \leq g_l + p_l \rightarrow \mu_{Z_l}(x) = 1 - \frac{Z_l(X) - g_l}{p_l} \quad (6)$$

$$if Z_l(X) \geq g_l + p_l \rightarrow \mu_{Z_l}(x) = 0 \quad (7)$$

The term $\mu_{Z_l}(x)$ indicates the desirability of the DM to solution X in terms of objective l . Then, the decision space is defined as the intersection of a membership function of objectives and system constraints as follows:

$$Z^* = \left\{ \bigcap_{l=1}^L Z_l^* \cap G \right\} (0 \leq \alpha \leq 1) \quad (8)$$

where, G is system constraints or feasible space. According to the extension principle, the membership function of Z is defined as follows:

$$\mu_Z(X) = \min\{\mu_{Z_l}(x)\} \quad (9)$$

Finally, the optimal solution, X^* , must maximize $\mu_Z(X)$ by solving the following equivalent crisp (c-MIGP) formulation of the f-MIGP.

3.2 The f-MIGP Model for Multi-Objective MMAL Sequencing Problem

The following solution procedure is employed to solve the f-MIGP for the proposed mixed-model assembly line sequencing problem:

Step 1: Construct the f-MIGP for the multi-objective mixed-model assembly line sequencing formulation.

Step 2: Solve the l^{th} objective function with an optimization technique, i.e. the lingo 8 software, and set g_l to the objective function value of the found minimum solution.

Step 3: Determine the values of the other objective functions of the obtained sequence in the previous step and set P_l = the maximum value among the obtained values - g_l .

Step 4: Repeat Step 2 and 3 for all the objective functions.

Step 5: Define the membership function of each fuzzy goal in the f-MIGP.

Step 6: Construct the equivalent crisp (c-MIGP) formulation of the f-MIGP.

Step 7: Solve the equivalent crisp (c-MIGP) formulation and obtain the decision level (Aspiration level α).

4 Experimental Results

4.1 Small-Sized Problems

The first experiment is carried out on a set of small-sized problems. The basic configuration of the problem set follows the sequencing problem found in Bard, et. al. ([1,2]). It must be noted that sequence-dependent setup costs occur only at the first station and the velocity of conveyor, v_c , is set to 1 for the sake of computational convenience. The nine MPS shown in Table 1 are tested. Table 2 represents the total obtained results for nine test problems.

Table 1. Problem sets

Problem	I	MPS	No. of feasible solutions	Launch interval
1	9	(4,3,2)	1260	6.2
2	9	(3,5,1)	504	6.6
3	10	(5,3,2)	2520	6.1
4	10	(4,4,2)	3150	6.3
5	10	(4,3,3)	4200	6.1
6	11	(4,6,1)	2310	6.6
7	11	(6,3,2)	4620	6.0
8	11	(5,3,3)	9240	6.0
9	12	(6,4,2)	13860	6.2

4.2 Large-Sized Problems

Another experiment is implemented on large-sized problems. In this experiment, the assembly time (t_{jm}), the length of each workstation (L_j), and the setup cost

Table 2. Obtained results from the lingo

Problem	Z_1^*	Z_2^*	Z_3^*	α	Optimal sequence
1	15.60	4.63	5	0.40	BAAAABBBCC
2	15.00	4.59	4	0.18	AABBBBBBCA
3	14.70	5.08	4	0.45	ABBBCCAAAA
4	16.71	4.26	5	0.40	AABBCCAABB
5	13.20	7.65	3	0.47	BBBCCCAAAA
6	15.40	6.63	7	0.17	BBBBBABCAAA
7	19.00	7.68	5	0.40	BBBAAAAAACCC
8	16.33	8.52	3	0.67	BBBCCCAAAAA
9	20.24	3.78	6	0.33	ABBCAAAAABBC

(c_{jmr}) are generated from uniform distributions of $U(4, 9)$, $U(12, 15)$, and $U(1, 3)$, respectively. The number of stations is fixed to 10 and the assembly line must produce five product models. Four MPSs shown in Table 3 are provided. Table 4 shows total obtained results for four test problems.

Table 3. Problem sets

Problem	I	MPS	No. of feasible solutions
1	20	(4,4,4,5,3)	2.44×10^{11}
2	20	(5,3,3,4,5)	1.95×10^{11}
3	20	(6,2,2,5,5)	5.87×10^{10}
4	30	(6,6,6,7,5)	1.18×10^{18}

Table 4. Obtained results from the lingo

Pr	Z_1^*	Z_2^*	Z_3^*	α	Optimal sequence
1	84.12	14.63	9	0.72	DDCCCCDAAAADDEEBBBBE
2	74.36	15.36	12	0.35	DDAAAADACCCDEEEEBBEE
3	78.15	10.42	15	0.18	ADCDDAAAADDEEEEBBDCA
4	129.36	12.58	26	0.10	DEEBCCDDAACDCCCBDAAADDEBBBBCEE

5 Conclusions

We have considered three objectives simultaneously minimizing the total utility work, total production rate variation, and total setup cost. Because of the natural conflict between these main objectives, the fuzzy goal programming can be an efficient approach for achieving a desirable solution from the decision maker (DM) point of view. This work has been done for the first time in solving a multi-objective mixed-model assembly line sequencing problem. We solve a comprehensive example and show that the proposed approach can determine the optimal sequence of products models with aim of maximizing the aspiration level and with respect to the given tolerance values.

References

1. Bard, J. F., Dar-El, E. M., Shtub, A. An analytic framework for sequencing mixed model. *International Journal of Production Research* 1992; 30: 35–48.
2. Bard, J. F., Shtub, A., Joshi, S. B. Sequencing mixed-model assembly lines to level parts usage and minimize the length. *International Journal of Production Research* 1994; 32: 2431–2454.
3. Hyun, C.J., Kim, Y., Kim, Y.K. A genetic algorithm for multiple objective sequencing problems in mixed model assembly lines. *Computers and Operations Research* 1998; 25 (7-8): 675–690
4. McMullen, P.R. JIT sequencing for mixed-model assembly lines with setups using tabu search *Production Planning and Control* 9 1998; (5): 504–510
5. McMullen, P.R. An efficient frontier approach to addressing JIT sequencing problems with setups via search heuristics. *Computers and Industrial Engineering* 2001a; 41: 335–353.
6. McMullen, P.R. A Kohonen self-organizing map approach to addressing a multiple objective, mixed-model JIT sequencing problem. *International Journal of Production Economics* 2001b; 72: 59–71.
7. McMullen, P.R. An ant colony optimization approach to addressing a JIT sequencing problem with multiple objectives. *Artificial Intelligence in Engineering* 2001c; 15: 309–317.
8. Miltenburg, J. Level schedules for mixed-model assembly lines in just-in-time production systems. *Management Science* 1989; 35 (2): 192–207.
9. Miltenburg, J., Steiner, G., Yeomans, S. A dynamic programming algorithm for scheduling mixed-model just-in-time production systems. *Mathematical Computation Modeling* 1990; 13: 57–66.
10. Monden, Y. *Toyota production system*. Institute if industrial engineers Press. Atlanta, GA: 1983.
11. Bellman, R.E., Zadeh, L.A. Decision making in a fuzzy environment, *Management Sciences* 1970; 17: 141–164.
12. Collette, Y., Siarry, P. *Multiobjective Optimization: Principles and Case Studies*. Springer, 2003.
13. Hannan, E.L. Linear programming with multiple goals, *Fuzzy Sets and Systems* 1981; 6: 235–164

On Asymptotically Optimal Algorithm for One Modification of Planar 3-dimensional Assignment Problem

Yury Glazkov

Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia
ygv@gorodok.net

Summary. In the paper the m -layer planar 3-dimensional assignment problem is considered. It is an NP-hard modification of well-known planar 3-dimensional assignment problem. Approximation algorithm, proposed by Gimadi and Korkishko is analysed. Its asymptotical optimality for a special class of random instances is proved.

1 Introduction

Let $C = (c_{ijk})$ be a $n \times n \times n$ 3-dimensional matrix with nonnegative elements. The problem of finding n transpositions $\pi^1, \dots, \pi^n$ with constraints $\pi^i(k) \neq \pi^j(k)$ for all $1 \leq i < j \leq n$ which minimize (or maximize) the sum $\sum_{i=1}^n \sum_{k=1}^n \pi^i(k)$ is a well-known planar 3-dimensional assignment problem (3-PAP).

If the first size of input matrix and the number of required transpositions is decreased to m , where $1 < m < n$, the problem is called m -layer planar 3-dimensional assignment problem (m-3-PAP). Though this modification is a simplification of the problem (one can interpret m-3-PAP as 3-PAP with $c_{ijk} = 0$ for $i > m$) it remains NP-hard problem (see f.e. [1]).

For 3-PAP an approximation algorithm was constructed by Kravtsov and Krachkovsky ([5]), but the proof of its asymptotical optimality for the problem with random instances failed (see [8]). For m-3-PAP an approximation algorithm was proposed in [3] and its asymptotical optimality was proved for the case when $m \sim \ln n$ and c_{ijk} has uniform distribution. Below this algorithm is exposed to more thorough analysis and its asymptotical optimality is proved for the case when $m \sim n^\theta$ ($\theta = \text{const}$) and c_{ijk} has minorized type distribution function. Also performance estimates of the algorithm are improved.

2 Approximation Algorithm A for m-3-PAP

Below we consider m-3-PAP formulated as above.

2.1 Algorithm A Description

Stage i , $i = 1, \dots, m$.

Consecutively for $j = 1, \dots, n - 2(i - 1)$ do:

1. Find $k_j = \operatorname{argmin}_{k=1, \dots, n} c_{ijk}$.

2. Set $\pi^i(j) = k_j$.

3. Forbid (set to infinity) elements $c_{ij'k_j}$ for $j' = j + 1, \dots, n$ and $c_{i'jk_j}$ for $i' = i + 1, \dots, m$.

Use Hopcroft-Karp algorithm proposed in [4] to solve (classical) assignment problem for the matrix formed by the rest $2(i - 1)$ lines ($j = n - 2(i - 1) + 1, \dots, n$) of i 'th layer of matrix C and construct the rest part of transposition π^i according to this solution.

2.2 Correctness and Running Time of Algorithm A

Lemma 1. *Algorithm A is correct for $m < n$ and its running time is $O(mn^2 + m^{7/2})$.*

Proof. Correctness of the algorithm is determined by existence of a feasible (without infinite elements) solution of (classical) assignment problem in the second part of each stage. As one can see from the description of the algorithm, on i 'th stage of it there will be at least $2(i - 1)$ unforbidden (finite) elements in each of lines (c_{ij}) for $j = n - 2(i - 1) + 1, \dots, n$. According to Ore theorem ([6]) it will be enough for the existence of required solution.

As the first part of each stage runs in time $O(n^2)$ and the second part (Hopcroft-Karp algorithm) runs with $O(m^{5/2})$ total running time of all m stages of algorithm A will be $O(mn^2 + m^{7/2})$.

The proof is complete.

3 Probabilistic Analysis of Algorithm A for Random Instances

Let us use the notations presented in [3]. Suppose algorithm A for solving optimization problem. By $F_A(I)$, $F^*(I)$ we denote the values of the objective function when solving an individual problem I by algorithm A and precisely respectively.

Algorithm A has performance estimates $(\varepsilon_n, \delta_n)$ in the class K_n of n -dimensional minimization problems if for all n the following inequality is satisfied:

$$\Pr\{F_A(I) > (1 + \varepsilon_n)F^*(I)\} \leq \delta_n \quad (1)$$

Here $\Pr\{J\}$ is the probability of a corresponding event J . ε_n is called a *performance ratio*, δ_n is a *fault probability* of algorithm A.

An algorithm is called *asymptotically optimal* in the problem class $K = \bigcup_{i=1}^{\infty} K_n$ if there exist its performance estimates $(\varepsilon_n, \delta_n)$ such that $\varepsilon_n \rightarrow 0$, $\delta_n \rightarrow 0$ as $n \rightarrow \infty$.

In the paper we consider the following type of random instances. Let all elements c_{ijk} of input matrix C be independent random variables distributed on segment $[a_n, b_n]$ where $b_n > a_n > 0$. Let $m \rightarrow \infty$ and $b_n/a_n \rightarrow \infty$ as $n \rightarrow \infty$.

3.1 General Probabilistic Analysis of Algorithm A

Let $\xi_{ij} = \pi^i(j)$ and $\xi'_{ij} = (\xi_{ij} - a_n)/(b_n - a_n)$ for $i = 1, \dots, m$ and $j = 1, \dots, n$

Consider left part of inequality (1). According to $F^*(x) \geq mna_n$ and $c_{ij\pi^i(j)} \leq b_n$

$$\begin{aligned} & \Pr\{F_A(x) > (1 + \varepsilon_A)F^*(x)\} \\ & \leq \Pr\left\{\sum_{i=1}^m \sum_{j=1}^n \xi'_{ij} > \varepsilon_A mn \frac{a_n}{b_n}\right\} \\ & \leq \Pr\left\{\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} \xi'_{ij} > \varepsilon_A mn \frac{a_n}{b_n} - m(m-1)\right\}. \end{aligned}$$

(as one can see we don't conduct probabilistic analysis of elements of transpositions choosed according to Hopcroft-Karp algorithm, just estimate it as in the worst case).

Now notice that ξ'_{ij} correspond to minimum of at least $n - j - i + 2$ elements of matrix C , so if η_{ij} ($i = 1, \dots, m, j = 1, \dots, n - 2(i - 1)$) are the similar random variables corresponding to exactly $n - j - i + 2$ elements of C then

$$\begin{aligned} & \Pr\left\{\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} \xi'_{ij} > \varepsilon_A mn \frac{a_n}{b_n} - m(m-1)\right\} \\ & \leq \Pr\left\{\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} \eta_{ij} > \varepsilon_A mn \frac{a_n}{b_n} - m(m-1)\right\}. \end{aligned} \quad (2)$$

In our analysis we will use the following theorem from [7].

Theorem 1. (Petrov's theorem) *Let $\eta_1, \dots, \eta_n$ - be independent random variables $S = \sum_{k=1}^n \eta_k$. If there exist some positive constants $g_1, \dots, g_n$, T such that*

$$Me^{t\eta_k} \leq \exp\left(\frac{1}{2} g_k t^2\right)$$

for all t , $0 \leq t \leq T$, then for $G = \sum_{k=1}^n g_k$

$$\Pr\{S > x\} \leq \begin{cases} \exp(-\frac{1}{2} x^2/G), & \text{if } 0 \leq x \leq GT, \\ \exp(-\frac{1}{2} Tx), & \text{if } x \geq GT. \end{cases}$$

It's easy to consider $\eta_{ij} - M\eta_{ij}$ instead of η_{ij} . If the first ones satisfy Petrov's theorem, then obtaining performance estimates for algorithm A can be reduced to bounding sums $\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} M\eta_{ij}$ and $\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} g_{n-i-j+2}$. So, this theorem and analysis above prove the main lemma of this subsection

Lemma 2. *Let random variables $\eta_{ij} - M\eta_{ij}$ ($i = 1, \dots, m, j = 1, \dots, n - 2(i - 1)$) satisfy Petrov's theorem with constants g_{ij} , T . Then algorithm A has performance estimates*

$$\varepsilon_A = \frac{m(m-1) + \overline{M} + T\overline{G}}{mn} \frac{b_n}{a_n}, \quad \delta_A = e^{-T\overline{G}/2},$$

where $\overline{M}$ and $\overline{G}$ - are any upper bounds of sums

$$\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} M\eta_{ij} \leq \overline{M}, \quad \sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} g_{n-i-j+2} \leq \overline{G}.$$

3.2 Performance Estimates of Algorithm A in the Case of Uniformly Distributed Instances

In this subsection we assume c_{ijk} be uniformly distributed on $[a_n, b_n]$. The following lemma proved in [2]

Lemma 3. *Let η_k be minimum of k independent uniformly distributed on $[0, 1]$ variables, and for $k = 1, \dots, n$ η_k are independent. Then for all t and k , $0 \leq t \leq 3$, $k = 1, \dots, n$, following inequality holds*

$$Me^{t(\eta_k - M\eta_k)} \leq \exp\left\{\frac{1}{2}g_k t^2\right\},$$

where

$$g_k = \begin{cases} 1/12 & \text{for } k = 1, \\ 11/36 & \text{for } k = 2, \\ \frac{11/5}{(k+1)^2} & \text{for } k = 3, \dots, n. \end{cases}$$

Lemma 4. *In the case of uniform distribution the following inequality holds*

$$\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} M\eta_{ij} \leq 2n \ln\left(\frac{n}{n-m}\right) + m \ln\left(\frac{n}{m}\right). \quad (3)$$

Lemma 5. *In the case of uniform distribution the following inequality holds*

$$\sum_{i=1}^m \sum_{j=1}^{n-2(i-1)} g_{n-i-j+2} \leq \frac{11}{5} \ln m + \frac{917}{180}. \quad (4)$$

Proof of lemmas (4) and (5) is based on evaluation of mathematical expectations and bounding sums with appropriate integrals.

Note, that bounds (3) and (4) are precise enough, but for better performance estimates of algorithm A we will use weakened bounds $\overline{M}$ and $\overline{G}$.

Theorem 2. *If*

$$b_n/a_n = \frac{n}{\varphi(n) \ln n}, \quad m = \Theta(\ln n),$$

where $\varphi(n) \rightarrow \infty$, $\varphi(n) = o\left(\frac{n}{\ln n}\right)$, then algorithm A is asymptotically optimal with performance estimates

$$\varepsilon_{A_1} = O\left(\frac{1}{\varphi(n)}\right), \quad \delta_{A_1} = n^{-\frac{3}{2} \ln n}.$$

Theorem 3. *If*

$$b_n/a_n = \frac{n^\theta}{\varphi(n)}, \quad m = \Theta\left(n^{1-\theta}\right),$$

where $\varphi(n) \rightarrow \infty$, $\varphi(n) = o(n^\theta)$, $\theta = \text{const}$, $0 < \theta < 1$, then algorithm A is asymptotically optimal with performance estimates

$$\varepsilon_{A_1} = O\left(\frac{1}{\varphi(n)}\right), \quad \delta_{A_1} = \exp\left(-\frac{3}{2}n^{2(1-\theta)}\right).$$

Proof of each theorem is based on using of weakened bounds (3) and (4) in lemma (2) under the conditions of the theorem.

Note, that all bounds imply large enough values of n . Constant $3/2$, in expression for δ_A , can be easily increased with negligible increase of ε_A .

3.3 The Case Minorized Type of Distribution Function

In this subsection we will show briefly that the performance estimates of algorithm A for some distribution function p of c_{ijk} remain correct for all distribution functions f minorized by p .

Lemma 6. *Let $\xi_1, \dots, \xi_k$ be independent random variables with distribution function $f(x)$, $\hat{f}(x)$ be distribution function of $\xi = \min_{i=1\dots k}(\xi_i)$, $\eta_1, \dots, \eta_k$ be independent random variables with distribution function $g(x)$, $\hat{g}(x)$ be the distribution function of $\eta = \min_{i=1\dots k}(\eta_i)$. Then for all x*

$$f(x) \leq g(x) \Rightarrow \hat{f}(x) \leq \hat{g}(x).$$

Lemma 7. *Let $p_\xi, p_\eta, p_\zeta, p_\chi$ be distribution functions of random variables ξ, η, ζ, χ respectively, ξ and ζ are independent, η and χ are independent. Then*

$$(\forall x \ p_\xi(x) \leq p_\eta(x)) \wedge (\forall y \ p_\zeta(y) \leq p_\chi(y)) \Rightarrow (\forall z \ p_{\xi+\zeta}(z) \leq p_{\eta+\chi}(z)).$$

Proof of lemmas (6) and (7) consists of using simple properties of minimum and sum of independent variables.

Theorem 4. *Let distribution functions $f(x)$ and $p(x)$ such that for all x $f(x) \geq p(x)$ and $(\varepsilon_A, \delta_A)$ are the performance estimates of algorithm A in the case of distribution function $p(x)$ then algorithm A has the same performance estimates in the case of distribution function $f(x)$.*

Proof of the theorem consists of applying lemmas (6) and (7) to the definition of performance estimates of algorithm A .

Corollary. Theorems 2 and 3 remain correct for all minorized types of distribution.

References

1. Frieze A (1983) Complexity of a 3-dimensional assignment problem. European J. Oper. Res. V 13, N 2: 161–164.
2. Gimadi E (2005) On some probability inequalities in some discrete optimization problems. // Operation Research Proceedings 2005, to appear in 2006.
3. Gimadi E, Korkishko N (2004) On some modifications of three index planar assignment problem // Discrete optimization methods in production and logistics. The second int. workshop (Omsk, July 20-27, 2004). Proc. DOM'2004, Omsk, 2004. P. 161-165.
4. Hopcroft J, Karp R (1973) An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. SIAM J. Comput. V. 2, N 4: 225–231.
5. Kravtsov M, Krachkovskii A (2001) O polinomionalnom algoritme nahozhdeniia asimptoticheski optimalnogo resheniia trehindeksnoi planarnoi problemi vibora. Zhurnal vychislitelnoi matematiki i matematicheskoi fiziki T. 41, ü 2: 342–345.
6. Ore O (1980) Teoriia grafov. Nauka, Moskva
7. Petrov V (1987) Predelnie teoremi dlia summ nezavisimih sluchainih velichin. Nauka, Moskva
8. Vozniuk I, Gimadi E, Fialtov M (2001) Asimptoticheski tochnii algoritm dlia resheniia zadachi razmesheniia s ogranichennimi obiemami proizvodstva. Diskretnii analiz i issledovanie operatsii Ser. 2. T. 8, ü 2:3–16.

A Multi-Objective Particle Swarm for a Mixed-Model Assembly Line Sequencing

Seyed Mohammed Mirghorbani, Masoud Rabbani,
Reza Tavakkoli-Moghaddam, and Alireza R. Rahimi-Vahed

Department of Industrial Engineering, University of Tehran, Iran
mohammad.mirghorbani@gmail.com

Summary. Mixed-model assembly line sequencing is one of the most important strategic problems in the field of production management. In this paper, three goals are considered for minimization; That is, total utility work, total production rate variation, and total setup cost. A hybrid multi-objective algorithm based on Particle Swarm Optimization (PSO) and Tabu Search (TS) is devised to solve the problem. The algorithm is then compared with three prominent multi-objective Genetic Algorithms and the results show the superiority of the proposed algorithm.

1 Introduction

A mixed-model assembly line is a type of production line capable of producing several small-sized production lots.

The structure of this paper is as follows. In section 2 considered problem is described. The Proposed algorithm will be treated in section 3. Experimental results will be given in section 4 and finally, the paper will be concluded in section 5.

2 Multi-Objective Sequencing Problem in MMAL

A MMAL considered in this paper is a conveyor system moving at a constant speed (v_c). The line is partitioned into J stations. It is assumed that the stations are all closed types. The worker moves downstream on the conveyor while performing his/her tasks to assemble a product. The design of the MMAL involves several issues such as determining operator schedules, product mix, and launch intervals. Early-start operator schedule is used in this paper (Hyun, et. al. [2]). In addition, the minimum part set (MPS) production, which is widely accepted in mixed model assembly lines, is also used in this paper. MPS is a vector representing a product mix, such that $(Q_1, Q_2, \dots, Q_M)$; where M is the total number of models. This strategy operates in a cyclical manner. The number of products produced in one cycle is given by $(I = \sum_{i=1}^M Q_i)$. The capacity of this line is limited with the maximum value, (C_{max}) , so that $(C_{max} \leq I)$. Third, the launch interval (γ) is set to $(\frac{T}{I \times J})$, in which T is the total operation time required to produce one cycle of MPS products.

The first objective considered is the minimization of the total utility work. The utility work is typically handled by the use of utility workers assisting the regular workers during the work overload. This objective is taken from Hyun, et al. [2]. The second objective is the minimization of the total production rate variation. One basic requirement of JIT systems is the continual and stable part supply. The objective can be achieved by matching demand with actual production. This objective is taken from Miltenberg [6]. The final objective is the minimization of the total setup cost. In a mixed Model Assembly Line, sequence-dependent setup costs constitute a considerable part of the total cost incurred to the system. Hence minimizing this cost will be highly beneficial for the system. This objective is taken from Hyun, et. al. [2].

3 The Proposed Multi-Objective Particle Swarm

Particle swarm was originated in 1995 by Kennedy and Eberhart [3] after studying the social behavior of birds.

3.1 Solution Representation

In addition to the typical job-to-position representation used in the context of MMAL problems, a real encoding is also used when applying the PSO. However, the real encoding does not represent a solution. To find the standard permutation of product models from real encoding, as many as the number of the first product models will be scheduled in the place of as many smallest numbers of the coordinates of the position vector of a particle. the next product models will be scheduled in the place of the next smallest values of the coordinates of the position vector, and so on. The assignment among each product type will be done arbitrarily. To obtain the job-to-position representation from real encoding as required, random numbers are generated and the smallest of them will be assigned to the first product model, the next smallest, to the second product model and so on.

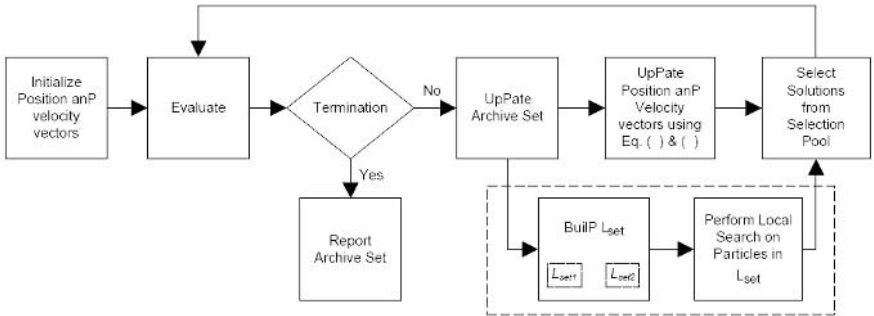


Fig. 1. Proposed Algorithm

3.2 Main Algorithm

In the initialization step, a new Elite Tabu Search (ETS) mechanism is used to construct the set of initial solutions to build a set of potentially diverse and high quality solutions. For this purpose, the approximation of the ideal point (where finding the exact point is not possible) is used and this approximation is updated at the end of each iteration. Initial velocity vectors are generated using the following equation:

$$\nu_{ij}^0 = \beta \times [r_1 \times \nu_{min} + (\nu_{max} - \nu_{min})] \quad (1)$$

Where $\beta = 1$ and r_1 is a random number between $[0 - 1]$. These values are restricted to some range, that is $[-4, +4]$. When updating the position and velocity vectors of the particles, the following equations are used:

$$v_{id}(k+1) = \chi[\omega_k v_{id}(k) + c_1 r_1 (p_{id}(k) - x_{id}(k)) + c_2 r_2 (p_{gd}(k) - x_{id}(k))] \quad (2)$$

$$x_{id}(k+1) = x_{id}(k) + v_{id}(k) \quad (3)$$

Where $p_{gd}(k)$ is the position vector for the solution in the archive set with the highest crowding distance. ω_k is the inertia weight, χ is the constriction factor, c_1 and c_2 are cognitive and social factors respectively and equal to 2, k is the iteration number and r_1 and r_2 are random numbers between $[0,1]$.

Local search is only performed on the set L_{set} . This set is a set of both diverse and high quality solutions. L_{set} , consists of two subsets, L_{set1} of maximum size b_1 and L_{set2} of size b_2 respectively. First, maximum b_1 solutions with the highest crowding distance will be selected from the archive set to construct L_{set1} . Then, those dominated and non-dominated solutions which do not exist in L_{set1} are identified and their minimum Euclidean distances to the solutions in L_{set} is computed. To build L_{set2} , b_2 solutions with the maximum of these minimum distances are selected and added to L_{set} .

The improvement method that we implemented is based on the parallel local search strategy. This parallel local search consists of two well-known operators, i.e. exchange procedure and insert procedure, which separately operate on each solution.

After updating velocity and position vectors of the particles with PSO equations, all of them along with the improved particles from the local search will be moved to the selection pool to be selected for the next iteration. Since the size of the pool in all likelihood will exceed swarm size, those particles with the lowest distance to the current approximation of the ideal point will be transferred to the next iteration. If some of the particles chosen for the next iteration are from the local search, velocity vectors are generated for them using Eq.(1) with $\beta = 0.5$. It should be noted that the coordinates of the approximation vector of the ideal point will be updated at the end of each iteration according to the best values found for each of the objective functions. Besides that, when updating the archive set, if a new non-dominated solution is found when the archive is full, if it is located outside the duplication area of its nearest solution from the archive set it will still enter the archive set. Duplication area of a non-dominated solution in the Pareto archive is defined as a bowl of center of the solution and of radius λ .

4 Experimental Results

The experiments are implemented in two folds: first, for the small-sized problems, the other for the large-sized ones. For both of these experiments, we consider the following assumptions:

General assumptions: (1) The velocity of the conveyor, v_c , is set to 1, and (2) Each experiment is repeated 15 times.

Multi-objective particle swarm's assumptions: (1) The values of c_1 and c_2 are fixed to 2, (2) the gbest model of Kennedy and Eberhart is deployed, (3) the value of χ is set to 0.73, and (4) The initial and final values for inertia weight which decreases linearly throughout the algorithm are set to 0.9 and 0.4 respectively.

Multi-objective genetic algorithms' assumptions: (1) The initial population is randomly generated, (2) The roulette wheel selection mechanism is used for PS-NC GA, (3) The binary tournament selection procedure is used for NSGA-II and SPEA-II, (4) The selection rate is set to 0.8, (5) The order crossover (OX) and inversion (IN) are used as crossover and mutation operators, and (6) The ratio of ox-crossover and inversion is set to 0.8 and 0.4, respectively.

To validate the reliability of the proposed MOPS, two comparison metrics are taken into account: The number of the Pareto solutions found by the algorithms, and Diversification metric, the results are then compared with those of PS-NC GA [2], NSGA-II [4] and SPEA-II [5],

4.1 Small-Sized Problems

The first experiment is carried out on a set of small-sized problems. The basic configuration of the problem set follows the sequencing problem in Bard, et. al. [1]. The four MPS's shown in Table 1 are tested. The last column refers to the launch interval for each problem instance.

Table 1. Problem sets for the small sized problems

Problem	I	MPS	No. of feasible solutions	Launch interval
1	10	(5,3,2)	2520	6.1
2	10	(4,4,2)	3150	6.3
3	10	(4,3,3)	4200	6.1
4	11	(4,6,1)	2310	6.6

the parameter setting is as follows: (1) The number of solutions at each iteration, N , is set to 50, (2) The algorithm is terminated after 40 iterations, (3) Since each objective function is linear ,the ideal point can be found. Therefore, the exact ideal point is used (4) The maximum size of L_{set1} , b_1 , is set to 15, and the size of L_{set2} , b_2 , is set to 10 (5) The maximum Pareto archive size, is fixed to 50, and (6) The radius of duplication area is set to 1,

4.2 Large-Sized Problems

For the large-sized problems, the number of workstations is set to 20 stations, length of each station j is fixed to be 1.5 , the number of models is set to $[0.25] + 1$, and

Table 2. Number of Pareto optimal solutions found in small-sized problems

Problem	No. of Pareto solutions	MOPS	PS-NC GA	NSGA-II	SPEA-II
1	22	19.21	17.6	18.2	18.07
2	37	27.41	22.93	22.27	20.13
3	40	32.65	24.93	25.73	24.33
4	44	36.07	32.13	31.2	31.33

setup costs are generated from a uniform distribution of $U(1, 3)$. in problem set I, assembly times T_{jm} are generated by a uniform distribution of $U(32, 48)$, and sequence-dependent setup costs occur only in three stations. In problem set II, I and VM are fixed to 30 and moderate, respectively.

The parameter setting in the large-sized problems is as follows: (1) The swarm size is increased to 200, (2) The algorithm is terminated after 50 iterations, (3) The maximum Pareto archive size is set to 200, (4) b1 and b2 are set to 50 and 25 (5) The radius of duplication area is fixed to 10.

Table 3. Two sets of large-sized test problems

Set1				Set2		
Problem	I	VM	MPS: VM	Problem	VA	PS
1	20	Low:[0,1]	(4,3,5,4,4):0.2	1	U(38,42)	1/20
2	20	Mod:[1,2]	(6,4,4,4,2):1.6	2	U(38,42)	2/20
3	20	High:[2,∞]	(3,4,6,6,1):3.6	3	U(38,42)	3/20
4	30	Low:[0,1]	(3,4,4,4,4,3,4):0.19	4	U(32,48)	1/20

Table 4. Number of non-dominated solutions found in large-sized problems

Problem	MOPS	PS-NC GA	NSGA-II	SPEA-II
I-1	174.21	152.21	160.42	149.77
I-2	159.13	147.45	149.74	142.55
I-3	160.54	150.06	155.65	139.06
I-4	173.42	161.23	158.6	160.38
Problem	MOPS	PS-NC GA	NSGA-II	SPEA-II
II-1	141.34	129.66	134.88	120.45
II-2	168.66	153.42	148.4	149.8
II-3	177.28	166.91	172.68	159.61
II-4	184.47	174.45	172.46	158.46

5 Conclusions

In this paper, a multi-objective sequencing algorithm was developed to solve a mixed-model assembly line sequencing problem in JIT environment. Different test problems were solved and the results obtained demonstrated the superiority of the proposed hybrid algorithm.

Table 5. Comparison of diversification metric for large-sized problems

Problem	MOPS	PS-NC GA	NSGA-II	SPEA-II
I-1	13.50	13.12	13.41	13.27
I-2	14.64	14.27	14.50	14.35
I-3	13.29	13.20	13.26	13.18
I-4	17.40	17.24	17.36	17.32

Problem	MOPS	PS-NC GA	NSGA-II	SPEA-II
II-1	17.19	16.72	17.14	16.56
II-2	18.05	17.54	17.85	17.71
II-3	18.40	18.25	18.30	18.21
II-4	17.61	17.28	17.34	17.30

References

1. Bard, J. F., Shtub, A., and Joshi, S. B. (1994) Sequencing mixed-model assembly lines to level parts usage and minimize the length. *International Journal of Production Research* 32: 2431-2454.
2. Hyun, C.J., Kim, Y., Kim, Y.K. (1998) A genetic algorithm for multiple objective sequencing problems in mixed model assembly lines. *Computers and Operations Research* 25 (7-8): 675-690
3. Kennedy, J., Eberhart, R.C. (1995) Particle swarm optimization, in: *Proceedings of the IEEE International Conference on Neural Networks 1995*; 4: 1942-1948.
4. K. Deb, S. Agrawal, A. Pratab et al., (2002) A fast elitist nondominated sorting genetic algorithm for multiobjective optimization: NSGA-II, *IEEE Trans. on Evolutionary Computation* 6 (4) 182-197.
5. E. Zitzler, M. Laumanns, L. Thiele, (2001) SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization, in: *Evolutionary methods for design, optimization and control with applications to industrial problems*, EUROGEN 2001, Athens, Greece
6. Miltenburg, J. (1989) Level schedules for mixed-model assembly lines in just-in-time production systems. *Management Science* 35 (2): 192-207.

FLOPC++

An Algebraic Modeling Language Embedded in C++

Tim Helge Hultberg

EUMETSAT, Darmstadt, Germany
`tim.hultberg@eumetsat.int`

Summary. FLOPC++ is an open source algebraic modeling language implemented as a C++ class library. It allows linear optimization problems to be modeled in a declarative style, similar to algebraic modeling languages, such as GAMS and AMPL, within a C++ program. The project is part of COmputational INfrastructure for Operations Research (COIN-OR) and uses its Open Solver Interface (OSI) to achieve solver independence.

1 Introduction

Algebraic modeling languages are important tools in Operations Research, their benefits are well known [1]: They allow a model representation which is readable by both humans and computers, using a syntax close to the notation used by most modelers. They automate the generation of model instances, freeing the modeler from the burden of generating the low level matrix format needed by the solver. In summary, they make implementation and maintenance of optimization models much easier.

However, traditional algebraic modeling languages do also have major weaknesses:

- Limited flexibility for integration with other software components
- Limited procedural support for algorithm development
- Limited model/program structuring facilities
- Limited extendibility

In order to mitigate these weaknesses, most algebraic modeling languages have incorporated procedural constructs, which allow the construction of algorithms around the optimization models. It has also become common to provide some sort of support for embedding the models in applications.

A natural alternative to extending the existing modeling languages is to implement modeling abstractions within an existing general-purpose object-oriented programming language, C++, in order to enable the “Formulation of Linear Optimization Problems in C++”. This is what FLOPC++ is all about.

The work on FLOPC++ began in year 2000 when the author had developed a production model for a Danish poultry butchery in GAMS and was faced with the practical problems of integrating the model in a GUI for collecting data and displaying the results. Discussing these problems with Søren Nielsen, he pointed me to his excellent paper [3], which was the main inspiration for the start of the development of FLOPC++. [2]

The first versions of FLOPC++ used the CPLEX callable library as its solver, but the need for supporting several independent solvers soon became apparent. Luckily, the simultaneous appearance of the Osi (Open Solver Interface) providing an abstract base class to a generic linear programming (LP) solver, along with derived classes for specific solvers, made this an easy task.

The algebraic representation of a model convey structural information that is hidden in the matrix representation required by the solver. Sometimes this information can be used to develop specialized solution algorithms using a general linear optimization solver for subproblems. Examples of such algorithms include decomposition, column generation and model specific cutting plane algorithms. The development of such algorithms is greatly facilitated by working in an environment where the algebraic representation of the model is available. This kind of algorithms can often be implemented in traditional modeling languages, but the implementations are mostly hopelessly inefficient because problems are passed to the solver via file interfaces and most importantly because slightly modified problems get regenerated from scratch. FLOPC++ offers the possibility to modify problem instances and warm start the solver from the last solution.

In Section 2 we give a brief introduction to the implementation of FLOPC++ without going into every aspect, before the conclusion in Section 3.

2 Implementation

Consider the following excerpt from the FLOPC++ model “transport.cpp” (the full model is one of the examples in the model library which comes with the distribution):

```
MP_set S(numS);           // Sources
MP_set D(numD);           // Destinations
MP_subset<2> Link(S,D);    // Transp. links (subset of S*D)
MP_data DEMAND(D);
MP_data SUPPLY(S);
MP_variable x(Link);
MP_constraint supply(S);
MP_constraint demand(D);

supply(S) = sum(Link(S,D), x(Link)) <= SUPPLY(S);
demand(D) = sum(Link(S,D), x(Link)) >= DEMAND(D);
```

We see that C++ classes are available to represent the basic algebraic modeling constructs: index sets, parameters, variables and constraints. There is also a class for representing dummy indices (MP_index), but since index sets can also be used as dummy indices, they are not needed for this particular model.

When the last statement is executed several data members of the `MP_constraint` object supply get instantiated. The dummy index “S” corresponding to the index set “S”, over which the constraint is defined, is stored (yes, “S” plays the role of both index set and dummy index). The type of the constraint, $\geq$, is stored and the value of two objects, left and right, of the type `MP_expression` get assigned.

The objects of the `MP_expression` class are symbolic representations of linear expressions (including dummy indices) and are implemented as reference counted pointers to `MP_expression_base`. In general `MP_expressions` are trees since the classes derived from `MP_expression_base` have one or two pointers to `MP_expression_base` as members, except for the two `TerminalExpression` classes `Expression_constant` and `VariableRef`. The complete `MP_expression_base` class hierarchy is shown below. The

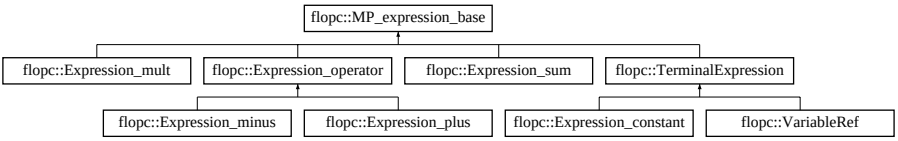


Fig. 1. `MP_expression_base` class hierarchy

`MP_expression` trees are build by a number of small operator overloaded functions and constructors as for example:

```

MP_expression operator+(const MP_expression& e1, const
MP_expression& e2)\{
    return new Expression_plus(e1, e2);}
  
```

and

```

MP_expression::MP_expression(const Constant &c) :
    Handle<MP_expression_base*>(new Expression_constant(c)) {}
  
```

The symbolic representation can then subsequently be used for generating the problem instances. In fact, it is this two step generation approach which makes the truly declarative notation of FLOPC++ possible in contrast to other modeling libraries such as Concert Technology from ILOG or LP Toolkit from Euro-decision.

The generation of the coefficient matrix entries takes place in the leaf nodes of the `MP_expressions` after informations (constant multipliers and domain) have been passed recursively down in the tree and combined from upper level nodes. When combining the domains, it is realized that $D * \text{Link}(S, D)$ is equal to $\text{Link}(S, D)$ so the complexity of generating the last constraint is of order $|\text{Link}(S, D)|$ and not $|D| * |\text{Link}(S, D)|$ as one might expect from a naive implementation.

As the very final step the coefficients matrix entries generated from the individual leaf node `TerminalExpressions` are assembled, i.e. eventual repeated entries are added together.

3 Conclusion

FLOPC++ is available from the home-page <https://projects.coin-or.org/FlopC++>, where detailed download and installation information can be found.

The main goal of FLOPC++ is to be as robust, efficient and easy to use for linear optimization as traditional algebraic modeling languages, such as AMPL and GAMS, while remaining lightweight and reaping the inherent benefits of being embedded in C++.

Acknowledgments

I would like to thank Philip Walton, Junction Solutions and JP Fasano, IBM who recently implemented several valuable improvements to FLOPC++ (Microsoft Visual C++ compatibility, doxygen comments, use of GNU autotools and full integration into Coin-Or, just to mention the most important) as well as the several other people who have reported bugs and suggested improvements over the last couple of years.

References

1. R. Fourer. Modeling languages versus matrix generators for linear programming. *ACM Transactions on Mathematical Software*, 1983.
2. T. Hultberg. *Topics in computational linear optimization*. PhD thesis, Department of Mathematical Modeling, Technical University of Denmark, 2000.
3. S. Nielsen. A C++ class library for mathematical programming. In *The impact of emerging technologies on computer science and operations research*. Kluwer, 1995.

Two-Machine No-Wait Flow Shop Scheduling Problem with Precedence Constraints

Saied Samie and Behrooz Karimi

Department of Industrial Engineering, Amirkabir University of Technology,
Tehran, Iran

saeedsamie@cic.aut.ac.ir

b.karimi@aut.ac.ir

Summary. In this paper, we consider two-machine permutation flow shop scheduling problem to minimize the makespan in which some of the jobs have to be processed with no-wait in process. We also consider precedence constraints which impose some jobs to be processed before or after some others. We propose a constructive algorithm to find a feasible solution. We also develop a Tabu Search algorithm to solve this problem. Our computational analysis indicates its good performance.

1 Introduction

In the classical two-machine flow shop scheduling problem, we are given a set of n jobs that have to be processed on two machines, A and B . The processing times of all jobs on two machines are given. Each machine processes one job at a time. Each job is processed first on machine A and then on machine B . No preemption is allowed. For a feasible schedule S the maximum completion time of all jobs on two machines is called the makespan and is denoted by $C_{max}(S)$. The resulting problem of minimizing the makespan is traditionally denoted by $F2||C_{max}$ and is known solvable in $O(n \log n)$ time.

In an important variation of $F2||C_{max}$, known as the flow shop with no-wait in process, it is assumed that every job starts on machine B exactly at the time of its completion on machine A . This problem, denoted by $F2|no-wait|C_{max}$, reduces to a special case of the traveling salesman problem (TSP) which can be solved in $O(n \log n)$ time [1, 2].

We consider a version of the two machine flow shop in which there are two types of jobs: regular and no-wait jobs [4]. Each regular job is processed as in problem $F2||C_{max}$, i.e., it starts on machine B no earlier than it is completed on machine A . Each no-wait job is processed as in problem $F2|no-wait|C_{max}$, i.e., it starts on machine B immediately after it is completed on machine A .

Extending standard scheduling notation, we denote the problem under consideration by $F2|reg+no-wait|C_{max}$. Let $N_R = \{J_1, \dots, J_{n_R}\}$ be the set of regular jobs, $|N_R| = n_R$ and $N_{NW} = \{I_1, \dots, I_{n_{NW}}\}$ be the set of no-wait jobs, $|N_{NW}| = n_{NW}$. We have $n = n_R + n_{NW}$ jobs to schedule. If either the number of regular jobs or the

number of the no-wait jobs is bounded by a constant q , the problem is denoted either by $F2|reg+no-wait, n_R \leq q|C_{max}$ or $F2|reg+no-wait, n_{NW} \leq q|C_{max}$, respectively.

For each of the basic problems $F2||C_{max}$ and $F2|no-wait|C_{max}$ there exists an optimal solution that is a permutation schedule in which each machine processes the jobs according to the same sequence. In general, this is not true for this extended model. In this model, the global optimal schedule may be achieved by allowing machines process the jobs in different sequences. But we restrict the search for an optimal solution to the class of permutation schedules, in which case we write $F2|reg+no-wait, pmu|C_{max}$ to denote the corresponding problem.

Problem $F2|reg+no-wait, pmu|C_{max}$ is *NP*-hard in the strong sense, provided that both n_R and n_{NW} are variable, i.e., part of the problem input. Also, it has been proved that the problem $F2|reg+no-wait, n_R \leq q, pmu|C_{max}$ is polynomially solvable and the problem $F2|reg+no-wait, n_{NW} \leq q, pmu|C_{max}$ is *NP*-hard in the ordinary sense even for $q=1$ [4].

Feasible schedules are further restricted by job precedence constraints given by partial order $\prec$, where $J_i \prec J_k$ means that job J_i must be processed before job J_k . Because of this new constraint, we shall denote the problem as $F2|reg+no-wait+prec, pmu|C_{max}$.

2 Generating a Feasible Schedule

We are given a set of precedence constraints in the form of $J_i \prec J_k$ which imposes job J_i to be processed before J_k in the permutation. To generate a feasible schedule, we first sort all jobs in an arbitrary order of their indices. This permutation may not be feasible because some constraint may not be satisfied. To change this unfeasible schedule to a feasible one, we use the following algorithm. This algorithm considers all constraints, one after the other and changes the position of jobs to satisfy the constraint. For example, if the constraint $J_i \prec J_k$ is being considered and the job J_i is placed in a position after job J_k , the algorithm moves the job J_i to be placed just before job J_k . But during this action, some other constraints that are considered earlier may not be longer satisfied. So the algorithm finds all jobs (J_l) between J_k and J_i that have precedence constraint with J_i in the form $J_l \prec J_i$ and moves them with job J_i to keep relating constraints satisfied. The algorithm is as below:

Step 1. Sort all jobs in an arbitrary order of their indices.

Step 2. Consider the first constraint. If this constraint is satisfied, go to step 6 and if not, call first job of constraint as *Pre* and the second job as *Suc* and go to next step.

Step 3. Find all jobs that are placed after *Suc* and meanwhile should be processed before *Pre* and put them in set K . If any job is added to set K , go to next step and if not, just add the *Pre* to set K and then go to step 5.

Step 4. Consider individually all the jobs which are added recently to the set K and find the jobs that are placed after *Suc* and meanwhile should be processed before the considered job. If there exists any new job, add it to the set K . Repeat this step for the last new added jobs in set K . If the set K remains unchanged, add the *Pre* to set K and go to the next step.

Step 5. Correct the order of jobs in set K according to their order in current permutation and insert this ordered sequence of jobs just before the job Suc and go to the next step.

Step 6. If there is any unchecked constraint, Check if this constraint is satisfied or not. If it is satisfied, repeat this step and if not, Call first job of constraint as Pre and the second job as Suc and go to step 3. If there exists no more constraint, the current permutation is a feasible one.

3 Heuristic Algorithm

Currently, Tabu Search is one of the most effective methods of the local search techniques designed to find a near-optimal solution of many combinatorial intractable optimization problems, such as the vast majority of scheduling problems. Already, TS has been applied to $Fm|no-wait|C_{max}$ [5]. We also propose a TS approach for our problem.

The components used in our algorithm have been proposed by Grabowski and Pempera [5], where they have applied on descending search and TS algorithms for no-wait flow shop problem.

3.1 Moves and Neighborhood

One of the main components of a local search algorithm is the definition of the move set that creates a neighborhood. Let $v = (x, y)$ be a pair of position in a permutation π and $x, y \in \{1, 2, \dots, n\}$, $x \neq y$. The pair $v = (x, y)$ defines a move in π . Among many types of moves considered in the literature, two of them appear prominently:

- (i) Insert move (I-move) removes the job $\pi(x)$ from its original position x and inserts it in a position y in $\pi(y)$.
- (ii) Swap move (S-move), in which the jobs $\pi(x)$ and $\pi(y)$; $x \neq y$ are interchanged in some positions x and y in π .

From the literature, it follows that S-moves are not better than I-moves to find good permutations [5]. Therefore, in our algorithm, we will consider only I-moves.

For I-moves, the neighborhood $N(Z, \pi)$ is generated by the move set $Z = \{(x, y) | x, y \in \{1, 2, \dots, n\} \text{ and } y \notin \{x, x-1\}\}$ of the cardinality $(n-1)^2$ (the condition $y \notin \{x, x-1\}$ has been added to avoid redundancy of moves). Among the moves in set Z , only those moves are allowed that performing them does not yield an infeasible sequence. So we redefine set Z as:

$$Z = \{(x, y) | x, y \in \{1, 2, \dots, n\} \text{ and } y \notin \{x, x-1\} \text{ and } \pi_{v=(x,y)} \text{ is feasible}\} \quad (1)$$

3.2 Search Process

The neighborhood of an initial basic permutation π is searched in the following manner. First, the “best” move $v^* \in Z$ that generates the permutation $\pi_{v^*} \in N(Z, \pi)$ with the lowest makespan is chosen, i.e. $C_{max}(\pi_{v^*}) = \min_{v \in Z} C_{max}(\pi_v)$. If $C_{max}(\pi_{v^*}) < C^*$ (where C^* is the best makespan found so far), then the move v^* is selected for the search process. Otherwise, i.e. if $C_{max}(\pi_{v^*}) \geq C^*$, then the set of unforbidden moves (UF) that do not have the tabu status, is defined:

$$UZ = \{v \in Z \mid \text{move } v \text{ is } UF\} \quad (2)$$

Next, the “best” move that generates the permutation $v^* \in UZ$ with the lowest makespan, i.e. $C_{max}(\pi_{v^*}) = \min_{v \in UZ} C_{max}(\pi_v)$, is chosen for the search. Choosing a non-better move is limited to be done for k times.

If the move v^* is selected, then a pair of jobs corresponding to the move v^* is added to the tabu list and the resulting permutation π_{v^*} is created. Next, this permutation becomes the new basic one, i.e. $\pi = \pi_{v^*}$ and algorithm restarts to next iteration. If all moves from Z are forbidden (a very rare case), i.e. if $UZ = \emptyset$, then the oldest element of tabu list is deleted and the search is repeated until a UF -move is found.

3.3 Tabu List and Tabu Status of Move

In our algorithm, we use the tabu list defined as a finite list (set) T with length $LengthT$, containing ordered pairs of jobs. The list is initiated by introducing $LengthT$ empty elements. If a move $v = (x, y) \in Z$ is performed on permutation π generating π_v , then the pair of jobs $(\pi(x), \pi(x+1))$ if $x < y$, or the pair $(\pi(x-1), \pi(x))$ if $x > y$, representing a precedence constraint, is added to T . Each time before adding a new element to T , we must remove the oldest one. With respect to a permutation π , a move $v = (x, y)$ is forbidden, i.e. it has *tabu* status, if $A(\pi(x)) \cap \{\pi(x+1), \pi(x+2), \dots, \pi(y)\} \neq \emptyset$ if $x < y$, and $B(\pi(x)) \cap \{\pi(y), \pi(y+1), \dots, \pi(x-1)\} \neq \emptyset$ otherwise, where

$$A(j) = \{i \in J \mid (j, i) \in T\} \quad (3)$$

$$B(j) = \{i \in J \mid (i, j) \in T\} \quad (4)$$

Set $A(j)$ (or set $B(j)$) indicates which jobs are to be processed after (or before) job j with respect to the current content of the tabu list T .

4 Computational Results

Computational experiments are conducted to test the performance of the heuristic approach. The algorithm is coded in MATLAB and run on a PC Celeron 2.26GHz. First, we develop some sample problems in small size and compare the solutions obtained from the algorithm with the optimal. At the second step, we develop some other sample problems in large size and compare the result of our tabu search algorithm with another metaheuristic algorithm.

In both steps, number of regular and no-wait jobs is determined by using a random number which implies the number of both types are not restricted. The processing time on machines A and B are generated from a uniform integer distribution. Another feature of our problem is precedence constraints. For each case at the second step, number of precedence constraints is considered about one third of number of total jobs. Table 1 shows the result of experiments in first case.

We can observe from table 1 that differences between the heuristic and the optimal solution are small. It is expected that as the problem size increases, these differences do not increase so much and the heuristic algorithm obtains near optimal solution. We also examine our algorithm in large size and compare its solution

Table 1. Comparison between heuristic and optimal in small size

Regular Jobs	No-Wait Jobs	Precedence Constraints	Optimal Solution	Heuristic Solution	Deviation (%)
3	2	3	32.589	32.589	0
1	4	4	44.650	44.650	0
2	4	3	31.673	31.905	0.732
3	3	4	31.830	34.035	6.927
4	3	3	42.068	42.068	0
1	6	4	36.616	36.616	0
2	5	4	37.521	41.497	10.596
6	2	3	40.684	40.684	0
4	4	4	57.513	57.513	0
1	7	5	39.522	39.522	0
2	7	4	53.795	55.359	2.097
4	5	5	46.489	49.073	5.583
6	3	5	53.215	53.215	0

Table 2. Comparison between two heuristics in large size

Regular Jobs	No-Wait Jobs	Precedence Constraints	Tabu Search	Simulated Annealing	Difference (TS-SA)
17	33	15	2806.3	2801.1	5.2
11	49	20	3346.5	3366.6	-20.1
40	30	25	3741.3	3745.7	-4.4
58	22	25	4399.9	4408.2	-8.3
23	67	30	4704.1	4746.5	-42.5
32	68	32	5252.0	5319.3	-67.3
11	99	35	5570.7	5613.3	-42.6

to another algorithm which is based on simulated annealing. Table 2 shows this comparison.

In all cases in table 2, tabu search algorithm produces better solutions compared to simulated annealing.

References

1. P.C. Gilmore, R.E. Gomory, Sequencing a one-state variable machine: a solvable case of the traveling salesman problem, *Oper. Res.* 12 (1964) 655679
2. P.C. Gilmore, E.L. Lawler, D.B. Shmoys, Well solved special cases, in: E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, D.B. Shmoys (Eds.), *The Traveling Salesman Problem. A Guided Tour of Combinatorial Optimization*, Wiley, Chichester, 1985, pp. 87143
3. S.M. Johnson, Optimal two- and three-stage production schedules with setup times included, *Naval Res. Logist. Quart.* 1 (1954) 6168
4. J.-L. Bouquard, J.-C. Billaut, M.A. Kubzin, V.A. Strusevich, Two-machine flowshop scheduling problems with no-wait jobs, *Operations Research Letters* 33 (2005) 255–262
5. Jozef Grabowski, Jaroslaw Pempera, Some local search algorithms for no-wait & flow-shop problem with makespan criterion, *Computers & Operations Research* 32 (2005) 21972212

A Multi-Commodity Flow Approach for the Design of the Last Mile in Real-World Fiber Optic Networks^{*}

Daniel Wagner¹, Günther R. Raidl², Ulrich Pferschy³, Petra Mutzel⁴, and Peter Bachhiesl¹

¹ Department of Telematics and Network Engineering, Carinthia University of Applied Sciences, Austria

`d.wagner@fh-kaernten.at`

`p.bachhiesl@fh-kaernten.at`

² Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria

`raidl@ads.tuwien.ac.at`

³ Institute of Statistics and Operations Research, University of Graz, Austria

`pferschy@uni-graz.at`

⁴ Chair for Algorithm Engineering, University of Dortmund, Germany

`petra.mutzel@cs.uni-dortmund.de`

Summary. We consider a generalization of the Steiner tree problem on graphs suitable for the design of the last mile in fiber optic networks and propose a multi commodity flow formulation for the exact solution of this problem. Some experimental results are discussed.

1 Introduction

We consider the problem of finding a most cost-efficient fiber optic network to connect given customer nodes to an existing network infrastructure.

Given is a connected undirected graph $G = (V, E, c, l, p, k_{\max})$ describing the topology of the surrounding area of given customer nodes. Each edge $e = (i, j) \in E$ represents a straight segment of Euclidean length $l_e \geq 0$ where a fiber optics cable might be installed with construction costs $c_e \geq 0$. The existing infrastructure is given as a subgraph $I = (V_I, E_I)$ of G . Furthermore the set of nodes V consists of the customer nodes C that shall be connected and spatial nodes S (possible Steiner nodes) resulting from the underlying spatial topology. The customer set C is the disjoint union of C_1 and C_2 , whereby customers C_1 require a single connection and customers C_2 need to be redundantly connected.

^{*} This work is supported by the Austrian Research Promotion Agency (FFG) under grant 811378 and the European Marie Curie RTN ADONET under grant 504438.

In real-world network design we have to consider several technical and practical requirements which make the problem more difficult. We model them by the following constraints.

- *Junction constraint*: Customer nodes of set C have to be connected to the existing infrastructure. Attaching new connections to the infrastructure is only allowed at predefined junction nodes $J \subseteq V_I$.
- *Biconnectivity constraint*: Customer nodes of set C_2 need to be redundantly connected to the existing infrastructure by two node-disjoint paths.
- *$k_{\max}$ -redundancy constraint*: Occasionally, the biconnectivity constraint for the nodes in the set C_2 may be relaxed in the sense that such a node may be connected to any biconnected (Steiner or customer) node v via a single path of maximum Euclidean length $k_{\max}(k)$.
- *Non-crossing constraint*: Cable routes are not allowed to cross each other in a geometric sense.

The *Operative Planning Task* (OPT) consists of finding a minimum-cost connected subgraph $G' = (V_{G'}, E_{G'})$ of G , $V_{G'} \subseteq V$, $E_{G'} \subseteq E$ that connects all customers so that the technical constraints are satisfied.

An extended variant is the *Strategic Simulation Task* (SST) that represents a generalization of the prize-collecting Steiner tree problem [5]. For each customer $c \in C$ a prize $p_c \geq 0$, i.e. an intended return on investment, is introduced. The problem is to find a subgraph G' of G that connects a subset of customers, minimizing

$$\sum_{e \in E_{G'}} c_e + \sum_{i \in C \setminus V_{G'}} p_i$$

so that the technical constraints are satisfied as in the OPT variant.

2 An Approach Based on Multi Commodity Flows

Multi commodity flow formulations are known to yield strong descriptions for many network design tasks including various constrained spanning and Steiner tree problems [1, 3, 6]. The basic idea is to send in G one commodity from the root node 0, which results from shrinking the infrastructure I into a single node, to each customer node $k \in C$ to be connected. This already satisfies the junction constraint.

To realize the biconnectivity constraint for each customer node $k \in C_2$, two different commodities f^k and g^k are sent in G from the root node 0 to the customer node k , which may not share any edges or have nodes other than the root 0 and k in common. With respect to the $k_{\max}$ -redundancy constraint we introduce an auxiliary commodity h^k in the $k_{\max}$ -neighborhood of customer k that indicates the segment where commodities f^k and g^k flow along the same path according to the definition of $k_{\max}$ -redundancy.

We define the following node, edge and arc sets.

- Arcs connecting the root with spatial nodes: $A_0 = \{(0, j) \in E \mid j \in S\}$
- Edges connecting two spatial nodes with respect to customer $k \in C$: $E_S(k) = \{(i, j) \mid i, j \in V \setminus \{0, k\}\}$
- Pairs of reversely directed arcs corresponding to the edges in $E_S(k)$: $A_S(k) = \{(i, j), (j, i) \mid (i, j) \in V \setminus \{0, k\}\}$.

- Arcs leading to customer node k : $A(k) = \{(i, k) \in E\}$
- Let $A'(k) = A_0 \cup A_S(k) \cup A(k)$ be the set of all arcs relevant for one customer node $k \in C$.
- With respect to the non-crossing constraint let K be the set of all pairs of edges that are crossing.
- Let $N(k)$ be the set of all nodes with a path to $k \in C_2$, of length not longer than $k_{\max}(k)$ and $B(k)$ be the set of all arcs $(i, j) \in A'(k)$, $i, j \in N(k)$.

The sets K , $N(k)$ and $B(k)$ can be efficiently determined in advance.

The following variables are used in the integer linear programming formulation.

- Variables $x_{ij} \in \{0, 1\}$, $\forall (i, j) \in E$, indicate whether edge (i, j) is used ($x_{ij} = 1$) or not ($x_{ij} = 0$).
- For each $k \in C$ and each $(i, j) \in A'(k)$, we define a variable $0 \leq f_{ij}^k \leq 1$. They represent the flow of commodity f^k associated to node $k \in C$ from node i to node j via edge (i, j) .
- For each $k \in C_2$ and each $(i, j) \in A'(k)$, we further define the variable $0 \leq g_{ij}^k \leq 1$. They represent the flow of a second commodity g^k for each node $k \in C_2$ in order to achieve the required redundancy.
- Variables $y_i \in \{0, 1\}$, $\forall i \in C$, indicate in the SST problem whether customer node i is connected or not. In the OPT problem, these variables are fixed to 1.
- Variables $0 \leq h_{ij}^k \leq 1$, $\forall (i, j) \in B(k)$, $\forall k \in C_2$, indicate the used edges (i, j) forming the branch (i.e., non-redundant path) leading to node k .

The formulation as an integer linear program is as follows:

Minimize

$$\sum_{(i,j) \in E} c_{ij} x_{ij} + \sum_{i \in C} p_i (1 - y_i) \quad (1)$$

subject to

$$\sum_{(i,j) \in A'(k)} f_{ij}^k - \sum_{(j,i) \in A'(k)} f_{ji}^k = \begin{cases} -y_k & \text{if } j = 0 \\ y_k & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \quad \forall k \in C, \forall j \in V \quad (2)$$

$$\sum_{(i,j) \in A'(k)} g_{ij}^k - \sum_{(j,i) \in A'(k)} g_{ji}^k = \begin{cases} -y_k & \text{if } j = 0 \\ y_k & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \quad \forall k \in C_2, \forall j \in V \quad (3)$$

$$f_{ij}^k \leq x_{ij} \quad \forall (i,j) \in A_0 \cup A(k), \forall k \in C \quad (4)$$

$$f_{ij}^k + f_{ji}^k \leq x_{ij} \quad \forall (i,j) \in E_S(k), \forall k \in C \quad (5)$$

$$g_{ij}^k \leq x_{ij} \quad \forall (i,j) \in A_0 \cup A(k), \forall k \in C_2 \quad (6)$$

$$g_{ij}^k + g_{ji}^k \leq x_{ij} \quad \forall (i,j) \in E_S(k), \forall k \in C_2 \quad (7)$$

$$f_{ij}^k + g_{ji}^k \leq x_{ij} \quad \forall (i,j) \in A_S(k), \forall k \in C_2 \quad (8)$$

$$f_{ij}^k + g_{ij}^k \leq x_{ij} \quad \forall (i,j) \in A'(k) \setminus B(k), \forall k \in C_2 \quad (9)$$

$$f_{ij}^k + g_{ij}^k - h_{ij}^k \leq x_{ij} \quad \forall (i,j) \in B(k), \forall k \in C_2 \quad (10)$$

$$h_{ij}^k \leq f_{ij}^k \quad \forall (i,j) \in B(k), \forall k \in C_2 \quad (11)$$

$$h_{ij}^k \leq g_{ij}^k \quad \forall (i,j) \in B(k), \forall k \in C_2 \quad (12)$$

$$\sum_{(i,j) \in B(k)} (f_{ij}^k + g_{ij}^k - h_{ij}^k) + \sum_{(i,j) \in A'(k) \setminus B(k)} (f_{ij}^k + g_{ij}^k) \leq 1 \quad \forall i \in V \setminus \{0,k\}, \forall k \in C_2 \quad (13)$$

$$\sum_{(i,j) \in B(k)} l_{ij} h_{ij}^k \leq k_{\max}(k) \quad \forall k \in C_2 \quad (14)$$

$$x_e + x_{e'} \leq 1 \quad \forall (e,e') \in K \quad (15)$$

The objective function (1) corresponds to the usual one in the prize-collecting Steiner tree problem [4]. Equalities (2) and (3) are the flow conservation constraints, in which we distinguish root node, connection objects and possible Steiner nodes. Constraints (4) to (8) relate the flow variables to x_{ij} and prevent reversely directed flows over the same edge. Constraints (9) ensure for each arc (i,j) outside of the $k_{\max}$ -neighborhood of node $k \in C_2$ that the flows of commodities f^k and g^k do not both use it.

Inequalities (10) force $h_{ij}^k = 1$ if $f_{ij}^k = 1 \wedge g_{ij}^k = 1$, i.e., both commodities associated to a $k \in C_2$ are routed over the same arc (i,j) . If $f_{ij}^k = 0 \vee g_{ij}^k = 0$, inequalities (11) and (12) force $h_{ij}^k = 0$. Inequalities (13) limit the outgoing flow of each node $i \in S$ with respect to commodities f^k and g^k to one except for the case when both commodities leave the node over the same edge being part of the branch ($h_{ij}^k = 1$). Constraints (14) limit for each node $k \in C_2$ the total length of all edges for which $h_{ij}^k = 1$ to $k_{\max}(k)$, i.e., the maximum branch lengths are enforced. Finally, inequalities (15) guarantee the non-crossing constraint.

The number of variables (16) and constraints (17) used by the formulation is as follows:

$$|E| + |C| + \sum_{k \in C} |A'(k)| + \sum_{k \in C_2} (|A'(k)| + |B(k)|) \quad (16)$$

$$\begin{aligned} & (|V| + |E|) (|C| + |C_2|) + \sum_{k \in C_2} (|A_S(k)| + |A'(k)| + 2|B(k)|) \\ & + |C_2| (|V| - 1) + |K| \end{aligned} \quad (17)$$

3 Experimental Results

We tested our approach on Intel Xeon 3.6GHz machines with 4GB memory using the general purpose ILP solver CPLEX 10.0.1 from ILOG. For each individual run

a time limit of 7200 seconds was used. The tests were performed with artificial grid-graph instances and instances constructed from real-world data from a German city [2].

The tests were run for both the OPT and SST problem variants with and without the non-crossing constraint (NCR) enabled. These different problem configurations were tested with (Table 2) and without (Table 1) $k_{\max}$ -redundancy. The tests with $k_{\max}$ -redundancy enabled used a $k_{\max}$ of 20 for the G0100-I{1,2,3} instances, 10 for G0400-I{1,2}, 30 for the ClgS-E instances, 100 for the ClgM-E instances and 150 for the ClgN1B-I{1,2} sets.

The column *%-Gap* shows the average gap between the best integer feasible solution and the best lower bound from the branch and bound tree, the column *Opt* shows the number of instances from a set that were solved to optimality.

Table 1. Results without $k_{\max}$ -redundancy.

Set	#	Dimension				OPT		OPT NCR		SST		SST NCR	
		V	E	C ₁	C ₂	Opt	%-Gap	Opt	%-Gap	Opt	%-Gap	Opt	%-Gap
G0100-I1	15	100	342	10	7	15	0.00	15	0.00	15	0.00	15	0.00
G0100-I2	15	100	342	8	4	15	0.00	15	0.00	15	0.00	15	0.00
G0100-I3	20	100	342	6	4	20	0.00	20	0.00	15	0.00	15	0.00
G0400-I1	15	400	1482	7	4	10	1.74	10	0.96	9	1.83	8	2.74
G0400-I2	15	400	1482	9	7	9	0.90	10	1.16	8	1.20	8	1.68
ClgS-E	25	190	377	4	2	25	0.00	25	0.00	25	0.00	25	0.00
ClgM-E	25	1757	3877	5	2	14	0.90	16	0.63	14	0.86	17	0.95
ClgN1B-I1	20	2804	3082	9	3	17	0.13	17	0.11	18	0.12	19	0.05
ClgN1B-I2	20	2804	3082	4	5	19	0.00	19	0.00	20	0.00	20	0.00
ClgN1E-I1	20	3867	8477	4	7	3	3.48	4	4.72	2	6.82	4	4.47
ClgN1E-I2	20	3867	8477	6	4	1	5.11	0	4.84	1	5.00	1	5.66
ClgN1E-I3	20	3867	8477	7	6	2	8.81	1	8.81	1	9.18	1	6.67

Table 2. Results with $k_{\max}$ -redundancy.

Set	#	Dimension				OPT		OPT NCR		SST		SST NCR	
		V	E	C ₁	C ₂	Opt	%-Gap	Opt	%-Gap	Opt	%-Gap	Opt	%-Gap
G0100-I1	15	100	342	10	7	13	0.56	13	0.78	13	0.75	13	0.75
G0100-I2	15	100	342	8	4	15	0.00	15	0.00	15	0.00	15	0.00
G0100-I3	20	100	342	6	4	20	0.00	20	0.00	20	0.00	20	0.00
G0400-I1	15	400	1482	7	4	1	6.76	1	7.18	1	7.25	2	7.63
G0400-I2	15	400	1482	9	7	0	8.75	1	8.24	0	8.35	0	8.75
ClgS-E	25	190	377	4	2	25	0.00	25	0.00	25	0.00	25	0.00
ClgM-E	25	1757	3877	5	2	0	4.26	1	7.93	0	3.86	1	4.35
ClgN1B-I1	20	2804	3082	9	3	4	2.97	4	3.44	3	3.26	3	3.74
ClgN1B-I2	20	2804	3082	4	5	3	7.34	2	6.55	2	8.47	3	6.41

This results show that at least smaller or rather sparse problem instances can be solved with our model within reasonable time. Some larger problem instances could be solved to provable optimality, too. Furthermore the results indicate that the $k_{\max}$ -redundancy constraint has a great impact on the solvability of the problem instances, although we used moderate numbers for $k_{\max}$ to limit the number of additional variables and constraints that have to be introduced.

4 Conclusions and Future Outlook

We presented a generalization of the Steiner tree problem in graphs which allows to make a tradeoff between survivability and costs for the connection of customers in access networks.

Furthermore we proposed a multi commodity flow formulation for this problem. Our experiments showed that although the $k_{\max}$ -redundancy constraint is a relaxation of the more strict biconnectivity constraint it is much harder to solve.

Currently, we are investigating a formulation based on directed cuts. Cut formulations have shown to be superior to flow formulations in the context of Steiner tree problems and prize-collecting Steiner tree problems. So we expect to gain improved results with this approach.

References

1. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice-Hall, 1993.
2. P. Bachhiesl. The OPT- and the SST-problems for real world access network design – basic definitions and test instances. Working Report Nr. 01/2005, Department of Telematics and Network Engineering, Carinthia Tech Institute, Austria, 2005.
3. H. A. Eiselt and C.-L. Sandblom. *Integer Programming and Network Models*. Springer, 2000.
4. M. X. Goemans and D. P. Williamson. *The Primal-Dual Method for Approximation Algorithms and its Application to Network Design Problems*, chapter 4, pages 144–191. PWS Publishing, 1997.
5. I. Ljubić, R. Weiskircher, U. Pferschy, G. W. Klau, P. Mutzel, and M. Fischetti. An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical Programming*, 105(2–3):427–449, February 2006.
6. M. Pióro and D. Medhi. *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Morgan Kaufmann Publishers, 2004.

On the Cycle Polytope of a Directed Graph and Its Relaxations

Egon Balas¹ and Rüdiger Stephan²

¹ Tepper School of Business, Carnegie Mellon University, U.S.A.*
eb17@andrew.cmu.edu

² TU Berlin, Germany
stephan@math.tu-berlin.de

1 Introduction

This paper continues the investigation of the cycle polytope of a directed graph begun by Balas and Oosten [2]. Given a digraph $G = (N, A)$ and the collection $\mathcal{C}$ of its simple directed cycles, the cycle polytope defined on G is $P_{\mathcal{C}} := \text{conv} \{\chi^C : C \in \mathcal{C}\}$, where χ^C is the incidence vector of C . According to the integer programming formulation given in [2], $P_{\mathcal{C}}$ is the convex hull of points $x \in \mathbb{R}$ satisfying

$$x(\delta^+(i)) - x(\delta^-(i)) = 0 \text{ for all } i \in N \quad (1)$$

$$x(\delta^+(i)) \leq 1 \text{ for all } i \in N \quad (2)$$

$$\begin{aligned} -x(S, N \setminus S) + x(\delta^+(i)) + x(\delta^+(j)) &\leq 1 \text{ for all } S \subseteq N, 2 \leq |S| \leq n-2, \\ i \in S, j \in N \setminus S \end{aligned} \quad (3)$$

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n x_{\pi(i)\pi(j)} \geq 1 \text{ for all permutations } \pi \text{ of } N \quad (4)$$

$$x_{ij} \in \{0, 1\} \text{ for all } (i, j) \in A \quad (5)$$

Here $\delta^+(k)$ and $\delta^-(k)$ denote the sets of arcs directed out of and into node k , respectively; for $F \subseteq A$, $x(F) = \sum (x_{ij} : (i, j) \in F)$; and for $S, T \subseteq N$, $x(S, T) := \sum (x_{ij} : i \in S, j \in T)$.

For the rest of this paper, we assume that the digraph G is complete. The dimension of $P_{\mathcal{C}}$ is $(n-1)^2$ (see [2]).

The inequalities of the above formulation, as well as the *nonnegativity constraints* $x_{ij} \geq 0$, were shown in [2] to be facet defining for $P_{\mathcal{C}}$.

Balas and Oosten [2] claimed that the only facet defining inequalities that cut off the origin, i.e. those inequalities $\alpha x \geq \alpha_0$ with $\alpha_0 > 0$, are the *linear ordering constraints* (4) and the inequalities equivalent to them. This claim is false; and

* Research supported by the National Science Foundation through grant #DMI-0098427 and by the Office of Naval Research through contract N00014-97-1-0196.

the main contribution of this paper is to describe the rich family of facet defining inequalities for P_C that cut off the origin and are not equivalent to any inequality (4).

The remainder of this paper is organized as follows: Section 2 examines the relationship between the cycle polytope and some of its relaxations: the cycle cone, the dominant of the cycle polytope, the upper cycle polyhedron. Section 3 deals with the polar relationship between cycles and linear orderings. Section 4 derives a characterization of all the facets of the upper cycle polyhedron. Finally, section 5 focuses on those facet defining inequalities that have coefficients in $\{0,1,2\}$.

2 Some Polyhedra Associated with P_C

The *dominant* of P_C , denoted by $\text{dmt}(P_C)$, is the Minkowski sum of the cycle polytope and the nonnegative orthant, i.e. $\text{dmt}(P_C) := P_C + \mathbb{R}_+^A$. Unlike P_C , its dominant is full-dimensional, i.e. $\dim(\text{dmt}(P_C)) = n(n-1)$.

The *upper cycle polyhedron* $\mathcal{U}_C(G)$ defined on G is $P_C + \text{cone}(C)$, where $\text{cone}(C)$ is the cone generated by its cycles.

Proposition 1.

$$\mathcal{U}_C = \text{dmt}(P_C) \cap \text{cone}(C).$$

Proof. See Balas and Stephan [3].

It follows from Proposition 1 that $\dim(\mathcal{U}_C) = (n-1)^2$, $P_C \subset \mathcal{U}_C \subset \text{dmt}(P_C)$, and $\text{dmt}(P_C) = \text{dmt}(\mathcal{U}_C)$.

Next we address the relationship between valid inequalities and facets of P_C , $\mathcal{U}_C$ and $\text{dmt}(P_C)$. First, from $P_C \subseteq \mathcal{U}_C \subseteq \text{dmt}(P_C)$ it follows that any inequality valid for $\text{dmt}(P_C)$ is also valid for $\mathcal{U}_C$ and P_C . Next, from the definition of the dominant, if $\alpha x \geq \alpha_0$ is valid for $\text{dmt}(P_C)$, then $\alpha_0 \geq 0$ and $\alpha_{ij} \geq 0$ for all $(i,j) \in A$. As to the relationship between P_C and its dominant, we will make use of a result of Balas and Fischetti [1, Theorem 8]. Applied to our polyhedron, it amounts to the following.

Theorem 1. *Let $\alpha x \geq \alpha_0$, with $\alpha_0 > 0$, be a valid inequality for $\text{dmt}(P_C)$ that defines a facet of P_C . Then $\alpha x \geq \alpha_0$ is facet inducing for $\text{dmt}(P_C)$ if and only if the subgraph of G induced by those arcs (i,j) such that $\alpha_{ij} = 0$ contains a spanning tree.*

From Proposition 1 it follows that, if $\alpha x \geq \alpha_0$ is valid for $\mathcal{U}_C$, then $\alpha x \geq \alpha_0$ is the positive combination of some inequality $\alpha' x \geq \alpha_0$ valid for $\text{dmt}(P_C)$ and some equations $x(\delta^+(i)) - x(\delta^-(i)) = 0$, $i \in N$; hence again $\alpha_0 \geq 0$, but it may happen that $\alpha_{ij} < 0$ for some (i,j) . However, the next theorem and the two following facts imply that any valid inequality for $\mathcal{U}_C$ (and P_C) that cuts off the origin has an equivalent form with all coefficients nonnegative.

Theorem 2. *An inequality $\alpha x \geq \alpha_0$ is valid (is facet defining) for $\mathcal{U}_C$ if and only if $\alpha_0 \geq 0$ and $\alpha x \geq \alpha_0$ is valid (is facet defining) for P_C .*

Proof. See Balas and Stephan [3].

An inequality $\alpha x \geq \alpha_0$ is said to be in h -rooted form (for some $h \in N$) if $\alpha_{ih} = 0$ for all $i \in N \setminus \{h\}$. Balas and Oosten [2] showed that any valid inequality for P_C can be brought to h -rooted form by adding to it the equations $-\alpha_{ih}(x(\delta^+(i)) - x(\delta^-(i))) = 0$ for all $i \in N \setminus \{h\}$ such that $\alpha_{ih} \neq 0$. Moreover, they showed that an inequality $\alpha x \geq \alpha_0$ which defines a facet of P_C , cuts off the origin, and is in h -rooted form, satisfies $\alpha_{ij} \geq 0$ and $\alpha_{ij} + \alpha_{ji} = \alpha_0$ for all $(i, j) \in A$. Both facts hold also for $\mathcal{U}_C$.

Note, two inequalities that define different facets of the dominant, may define a single facet of P_C or $\mathcal{U}_C$. It is also possible for an inequality, or a family of inequalities, to be facet defining for $\text{dmt}(P_C)$ but not for P_C or $\mathcal{U}_C$. An example will be given in section 5.

3 Cycles and Permutations

The presence of the linear ordering constraints (1) in the description of the cycle polytope suggests some connection between cycles and linear orderings or permutations. In fact, it expresses the condition that every directed cycle of G and every transitive tournament in G corresponding to a linear ordering of its nodes, must have at least one arc in common. The linear ordering polytope P_{LO} , also known as the transitive tournament polytope P_{TT} , can be defined as the convex hull of integer solutions to the system

$$\begin{aligned} y_{ij} + y_{ji} &= 1 \text{ for all pairs of nodes } i, j \\ y_{ij} + y_{jk} + y_{ki} &\geq 1 \text{ for all distinct } i, j, k \in N \\ y_{ij} &\geq 0 \text{ for all } (i, j) \in A. \end{aligned} \tag{6}$$

While the integer solutions of (6) are precisely the transitive tournaments (linear orderings), (6) also has fractional extreme solutions, and the polytope defined by (6), i.e. the linear programming relaxation of P_{LO} , is known as the *generalized transitive tournament polytope*, P_{GTT} [4], [5].

Recall that for every polyhedron $P \subseteq \mathbb{R}^q$, the *blocking polyhedron*, or *blocker* of P is defined as $\text{bl}(P) := \{y \in \mathbb{R}_+^q : y^T x \geq a \text{ for all } x \in P\}$. A polyhedron $P \subseteq \mathbb{R}^q$ is said to be of the *blocking type* if there exist vectors $c_1, \dots, c_p$ in $\mathbb{R}^q$ such that $P := \text{conv}(c_1, \dots, c_p) + \mathbb{R}_+^q$, and blocking theory says [6] that if P is of the blocking type, then the facets of $\text{bl}(P)$ are in 1-1 correspondence with the vertices of P , and vice versa.

Note that the dominant of any polytope is of the blocking type. Brualdi and Hwang [5, Theorem 2.2] showed that $\text{bl}(P_{GTT}) = \text{dmt}(P_C)$ which implies that the facets of $\text{dmt}(P_C)$ are in 1-1 correspondence with the vertices of the polytope P_{GTT} defined by (6), a system of $\frac{3}{2}n(n-1) + 2\binom{n}{3}$ linear inequalities in $n(n-1)$ variables, where $n = |N|$. Hence the separation problem for $\text{dmt}(P_C)$, i.e. the problem of finding a facet that cuts off a given point, or showing that no such facet exists, amounts to minimizing a linear function over (6), a task executable in polynomial time.

Having thus found an efficient way to identify the facets of $\text{dom}(P_C)$, our next task is to use the results of section 2 to obtain from these facets, the corresponding facets of $\mathcal{U}_C$ and P_C .

4 All the Facets of $\mathcal{U}_C$

In the next theorem we establish a connection between the dominant of the cycle polytope defined on the complete digraph with n nodes, and the upper cycle polyhedron defined on the complete digraph with $n+1$ nodes. To mark these differences between the orders of the respective digraphs, we will denote by P_C^n and $\mathcal{U}_C^n$ the cycle polytope and the upper cycle polyhedron defined on the digraph with n nodes. Also, P_{GTT}^n denotes P_{GTT} defined on a graph with n nodes.

Theorem 3. *Consider the inequalities*

$$\alpha x \geq \alpha_0 \quad (7)$$

$$\alpha x + \alpha_0(x(\delta^+(n+1))) \geq \alpha_0 \quad (8)$$

where $\alpha \in \mathbb{Z}^A$ and $\alpha_0 > 0$ is integer.

The following statements are equivalent:

- (a) $\frac{1}{\alpha_0}\alpha$ is a vertex of P_{GTT}^n
- (b) (7) defines a facet of $\text{dmt}(P_C^n)$
- (c) (8) defines a facet of P_C^{n+1} and of $\mathcal{U}_C^{n+1}$ with $\alpha \geq 0$
- (d) (8) defines a facet of $\text{dmt}(P_C^{n+1})$

Proof. See Balas and Stephan [3].

Theorem 3 establishes a 1-1 correspondence between the vertices of P_{GTT}^n and those facets of $\mathcal{U}_C^{n+1}$ and P_C^{n+1} defined by $(n+1)$ -rooted inequalities. The procedure that takes (7) into (8) will be referred to as *node lifting*. Since every nontrivial facet defining inequality is equivalent for P_C^{n+1} and $\mathcal{U}_C^{n+1}$ to a $(n+1)$ -rooted inequality, all the nontrivial facet inducing inequalities for $\mathcal{U}_C^{n+1}$, i.e. all those facet inducing inequalities for P_C^{n+1} that cut off the origin, are accounted for in this way. On the other hand, not all facet inducing inequalities for $\text{dmt}(P_C^{n+1})$ are accounted for by the inequalities (8), because there may be facet inducing inequalities of type (7) for $\text{dmt}(P_C^{n+1})$ that cannot be obtained by node lifting an inequality for $\text{dmt}(P_C^n)$.

Note that while (α, α_0) defines a unique vertex $\frac{1}{\alpha_0}\alpha$ of P_{GTT}^n and a unique facet inducing inequality (7) for $\text{dmt}(P_C^n)$, the inequality (8) that induces the facet F^{n+1} of $\mathcal{U}_C^{n+1}$ (P_C^{n+1}) is not unique. The same facet F^{n+1} can be defined by any inequality equivalent to (8), i.e. obtained from (8) by adding a combination of the flow conservation equations (1). In particular, for each node $v \in \{1, \dots, n\}$, we obtain an inequality $\alpha'x + \alpha_0x(\delta^+(v)) \geq \alpha_0$ that is equivalent to (8) with respect to P_C^{n+1} and $\mathcal{U}_C^{n+1}$ but defines a facet of $\text{dmt}(P_C^{n+1})$ different from the one defined by (8). Deleting the node v then yields a valid inequality $\alpha'x \geq \alpha_0$ for $\text{dmt}(P_C^n)$ of the form (7), but with $\alpha' \neq \alpha$, in fact α' can be obtained from α through a permutation of its components. This inequality defines a facet of $\text{dom}(P_C^n)$ different from the one defined by (7). Furthermore, $\frac{1}{\alpha_0}\alpha'$ is a vertex of P_{GTT}^n , different from the vertex $\frac{1}{\alpha_0}\alpha$. There are other non-rooted inequalities equivalent to (8), i.e. defining the same facet F^{n+1} for P_C^{n+1} and $\mathcal{U}_C^{n+1}$, with some negative coefficients; but these are not facet inducing for $\text{dmt}(P_C^{n+1})$.

Theorem 3 makes it possible to start with some inequality $\alpha x \geq \alpha_0 \geq 1$, facet defining for $\text{dmt}(P_C^n)$, and derive facet defining inequalities with righthand side

α_0 for $P_C^m, \mathcal{U}_C^m$ and $\text{dmt}(P_C^m)$ for all $m \geq n+1$ by applying recursively the node lifting procedure implied by the theorem. The converse of this statement is not true, however, since the reverse procedure cannot be used recursively: given a facet defining inequality $\alpha x \geq \alpha_0 \geq 1$ for P_C^n , one can bring it to rooted form and derive from it, by deleting the root node, a facet defining inequality $\alpha' \geq \alpha_0$ for $\text{dmt}(P_C^{n-1})$, but one cannot derive anything for P_C^m or $\text{dmt}(P_C^m)$ for any $m \leq n-2$, unless the inequality $\alpha' x \geq \alpha_0$ happens to be facet defining for P_C^{n-1} , which is not the case in general.

The vertices of P_{GTT} have been studied in the 90's by several researchers. Brualdi and Hwang [5] investigated the half-integer vertices, i.e. vertices with components in $\{0, \frac{1}{2}, 1\}$. Borobia [4] gave a list of all half-integral vertices of P_{GTT}^n for $n = 6$ and 7. Already for $n = 8$, P_{GTT}^n has fractional vertices with components in $\{1, \frac{1}{2}, \dots, \frac{1}{k}\}$ for $k \in \mathbb{Z}_+, k \geq 3$ (see Nutov and Penn [7]).

5 Facet Defining Inequalities with Coefficients in $\{0,1,2\}$

The only nontrivial facet defining inequalities for $P_C, \mathcal{U}_C$ and $\text{dmt}(P_C)$ of the form $\alpha x \geq \alpha_0$ with $\alpha_0 = 1$ and $\alpha \geq 0$ are the linear ordering inequalities (4). They correspond to the integer vertices of P_{GTT} .

Next we turn to inequalities $\alpha x \geq \alpha_0$ with $\alpha_0 = 2$, $\alpha \geq 0$. We start with the *necklace inequality*.

$$2x(T) + x(A \setminus T \cup \tilde{T}) \geq 2, \quad (9)$$

where A is the arc set of the complete digraph $G = (N, A)$ with $|N| = n$ even, T is an alternating traveling salesman tour in G , i.e. one whose arcs alternate in their direction of traversal, and $\tilde{T}$ is the reverse of T .

Theorem 4. *For the complete digraph $G = (N, A)$ with $|N| = n \geq 6$ and even, the inequality (9) is valid and facet inducing for $\text{dmt}(P_C^n)$, but is not facet inducing for P_C^n or $\mathcal{U}_C^n$.*

Proof. See Balas and Stephan [3].

Given a facet defining inequality (9) for $\text{dmt}(P_C^n)$ with n even, say $\alpha x \geq 2$, we can apply to it the node lifting procedure of Theorem 3 to obtain an inequality $\alpha x + 2x(\delta^+(n+1)) \geq 2$, facet defining for $\mathcal{U}_C^{n+1}$ and P_C^{n+1} . We can then choose an arbitrary node $i \neq n+1$, and bring this inequality to i -rooted form. The resulting inequality, $\alpha' x + 2x(\delta^+(i)) \geq 2$, is still facet defining for $\mathcal{U}_C^{n+1}(P_C^{n+1})$, and if we now delete node i , we obtain another facet defining inequality for $\text{dmt}(P_C^n)$, $\alpha' x \geq 2$. This inequality, derived from the necklace inequality (9) through node lifting, re-rooting and node deletion has $\alpha'_{ij} = 2$ and $\alpha'_{ji} = 0$ for $(i, j) = (1, 2), (1, n)$, and (i, j) contained in the transitive closure of the two directed paths $P_1 := \{(k, k-1), (k-1, k-2), \dots, (4, 3), (3, 2)\}$ and $p_2 := \{(k, k+1), (k+1, k+2), \dots, (n-2, n-1), (n-1, n)\}$, where $k = \frac{n}{2} + 1$; and $\alpha'_{ij} = \alpha'_{ji} = 1$ for all remaining arcs. We will call it a *pendant inequality*.

Corollary 1. *The pendant inequalities induce facets of $\text{dmt}(P_C^n)$ for $n \geq 6$ and even.*

References

1. E. Balas and M. Fischetti, "On the Monotonization of Polyhedra," *Mathematical Programming* 78 (1997) 59-84.
2. E. Balas and M. Oosten, "On the cycle polytope of a directed graph," *Networks* 36:1 (2000) 34-46.
3. E. Balas and R. Stephan, "On the cycle polytope of a directed graph and its relaxations", MSRR# 675, Tepper School, Carnegie Mellon University, August 2006.
4. A. Borobia, " $(0, \frac{1}{2}, 1)$ matrices which are extreme points of the generalized transitive tournament polytope," *Linear Algebra Applications* 220 (1995) 97-110.
5. R.A. Brualdi and G.-S Hwang, "Generalized transitive tournaments and doubly stochastic matrices," *Linear Algebra Applications* 172 (1992) 151-168.
6. D.R. Fulkerson, "Blocking and anti-blocking pairs of polyhedra," *Mathematical Programming* 1 (1971) 127-136.
7. Z. Nutov and M. Penn, "On non- $\{0, \frac{1}{2}, 1\}$ extreme points of the generalized transitive tournament polytope, *Linear Algebra Applications* 233 (1996) 149-159.

Modelling Some Robust Design Problems via Conic Optimization

Diah Chaerani and Cornelis Roos

Department of Software Technology, Faculty of Electrical Engineering, Mathematics and Computer Sciences, Delft University of Technology, The Netherlands.

`d.chaerani@ewi.tudelft.nl`

`c.roos@ewi.tudelft.nl`

Summary. In this paper, we deal with modelling robust design problems via conic optimization. A robust design problem deals with finding a robust optimal solution of an uncertain design problem. The uncertain data is assumed to belong to a so-called uncertainty set $\mathcal{U}$. Uncertainty means that the data is not known exactly at the time when the solution has to be determined. In order to find a robust optimal solution, we use the robust optimization (RO) methodology of Ben-Tal and Nemirovskii. We demonstrate this on the robust shortest path problem (RSPP), the robust maximum flow problem (RMFP) and the robust resistance network topology design (RNTD) problem.

1 Introduction

Conic optimization (CO) is a very useful optimization technique that concerns the problem of minimizing a linear objective function over the intersection of an affine set and a convex cone. The general form of a conic optimization problem is as follows:

$$\min_{x \in \mathbf{R}^n} \{c^T x : Ax - b \in \mathcal{K}\}. \quad (\text{CP})$$

The objective function is $c^T x$, with objective vector $c \in \mathbf{R}^n$. Furthermore, $Ax - b$ represents an affine function from $\mathbf{R}^n$ to $\mathbf{R}^m$, $\mathcal{K}$ denotes a convex cone in $\mathbf{R}^m$ and the constraint matrix A is of size $m \times n$. The importance of conic optimization is due to the fact that it can model many nonlinear optimization problems. Furthermore, conic optimization problems can be solved efficiently using interior-point methods since the theory of self-concordant barriers developed by Nesterov and Nemirovskii [5] provides an algorithm with polynomial complexity.

We demonstrate how to use CO models on three typical problems, i.e., the robust shortest path problem (RSPP), the robust maximum flow problem (RMFP) and the robust resistance network topology design (RNTD) problem.

The paper is organized as follows. In Section 2 we present a short description on modelling robust design problems. The RSPP, the RMFP and the RNTD are presented in Section 3.1, 3.2 and 3.3, respectively. Some concluding remarks are presented in Section 4.

2 Modelling Robust Design Problems via CO

A robust design problem deals with finding a robust optimal solution of an uncertain design problem. The robust optimal solutions of these problems are obtained using the robust optimization (RO) methodology of Ben-Tal and Nemirovskii [1]. In this methodology, it is assumed that the data vector ζ belongs to an uncertainty set $\mathcal{U}$. Thus, an uncertain optimization problem can be expressed as follows.

$$\min\{f_0(x, \zeta) : f_i(x, \zeta) \leq 0, \quad i = 1, \dots, m\}, \quad \zeta \in \mathcal{U}. \quad (1)$$

One associates with the uncertain problem (1) its so-called robust counterpart which is given by

$$\min\{t : f_0(x, \zeta) \leq t, \quad f_i(x, \zeta) \leq 0, \quad i = 1, \dots, m, \quad \forall \zeta \in \mathcal{U}\}. \quad (2)$$

Obviously, this robust counterpart of (1) represents a worst-case oriented approach: a pair of solutions $(t; x)$ is feasible only if x satisfies the constraints in (1) for all possible values of ζ . The optimal solutions of (2) are called robust optimal solutions. The robust counterpart (2) is an optimization problem with usually infinitely many constraints, depending on the uncertainty set $\mathcal{U}$, thus this problem may be very hard to solve. This means that only if $\mathcal{U}$ is chosen suitably, the problem (2) can be solved efficiently.

We consider the uncertain optimization problem in conic form:

$$\min_{x \in \mathbf{R}^n} \{c^T x : A_i x - b_i \in \mathcal{K}_i, i = 1, \dots, m\}, \quad (c, \{A_i, b_i\}_{i=1}^m) \in \mathcal{U} \quad (3)$$

where for every i , $\mathcal{K}_i$ is a convex cone and where $\mathcal{U}$ is the uncertainty set. We restrict ourselves to the cases when each $\mathcal{K}_i$ is either the nonnegative orthant, or the Lorentz (second-order or ice cream) cone, or the positive semidefinite cone. Thus the robust counterpart to (3) can be expressed in the following CO model:

$$\min \left\{ t : c^T x - t \leq 0, A_i x - b_i \in \mathcal{K}_i, i = 1, \dots, m, \forall (c, \{A_i, b_i\}_{i=1}^m) \in \mathcal{U} \right\}. \quad (4)$$

3 Applications on Some Robust Design Problems

In this section we present the CO models of the three robust design problems mentioned before, i.e., the RSPP, the RMFP and the RNTD.

3.1 Robust Shortest Path Problem (RSPP)

Based on [3], in this section we discuss the problem of finding the robust shortest path. For a given network $G = (\mathcal{V}, \mathcal{A})$ with node set $\mathcal{V}$ and arc set $\mathcal{A}$, an instance of the SPP consists of finding a directed path from a single node s to a single node t with a minimal total length with respect to a given length function $c : \mathcal{A} \rightarrow \mathbf{R}_+$. A binary optimization model for the shortest path problem is

$$\min\{c^T x : Ax = b, x \in \{0,1\}^{\mathcal{A}}\}, \quad (5)$$

where A is the node-arc incidence matrix, b is a vector in $\mathbf{R}^{\mathcal{V}}$ with entries $b_s = 1$, $b_t = -1$ and $b_j = 0$ for all $j \in \mathcal{V} \setminus \{s, t\}$, and c is the vector of arc lengths. In the uncertain SPP, we assume that A and b are fixed. The uncertainty is only in the vector c , so $\mathcal{U}$ will denote the set of all possible realizations of c . Thus, the uncertain SPP is given by

$$\min_x \left\{ c^T x : Ax = b, x \in \{0,1\}^{\mathcal{A}}, c \in \mathcal{U} \right\}, \quad (6)$$

and the corresponding robust counterpart is

$$\min \left\{ \ell : \ell \geq c^T x, Ax = b, x \in \{0,1\}^{\mathcal{A}}, \forall c \in \mathcal{U} \right\}. \quad (7)$$

Using the set $\mathcal{P} := \{x : Ax = b, x \in \{0,1\}^{\mathcal{A}}\}$, which consists of the incidence vectors of all $s - t$ paths, we can reformulate (7) as

$$\ell^* = \min_{x \in \mathcal{P}} \left\{ \max_{c \in \mathcal{U}} c^T x \right\}, \quad (8)$$

with ℓ^* denoting the optimal value of (7). A path with worst case length ℓ^* is called a robust shortest path (RSP) and the problem of finding a RSP is called a robust shortest path problem (RSPP). Of course, the robust counterpart (7) depends on how we choose the uncertainty set $\mathcal{U}$. We consider two different uncertainty sets, namely boxes and ellipsoidal sets. The results are presented in the following theorem.

Theorem 1. *In case of the SPP with box uncertainty we have*

$$\mathcal{U} = \{c : l \leq c \leq u\}, \quad \text{where } l, u \in \mathbf{R}^{\mathcal{A}} \quad \text{with } l \leq u. \quad (9)$$

The robust counterpart of the uncertain SPP is a SPP with the original arc lengths replaced by their upper bounds

$$\ell^* = \min_{x \in \mathcal{P}} \left\{ \max_{c \in \mathcal{U}} c^T x \right\} = \min_{x \in \mathcal{P}} u^T x. \quad (10)$$

Thus, every sub path of the robust shortest path is a robust shortest path.

Theorem 2. *In case of the SPP with ellipsoidal uncertainty we have*

$$\mathcal{U} = \{c : c = c^n + \mathcal{Q}w, \|w\| \leq 1\}, \quad (11)$$

where c^n is the vector of nominal arc lengths and $\mathcal{Q}$ is a fixed matrix of size $|\mathcal{A}| \times p$ and $w \in \mathbf{R}^p$ for some p . The robust counterpart of the uncertain SPP is a conic quadratic optimization (CQO) problem with binary variables.

$$\min \{ \ell_x : \ell_x \geq (c^n)^T x + \|\mathcal{Q}^T x\|, Ax = b, x \in \{0,1\}^{\mathcal{A}} \}. \quad (12)$$

In this case a subpath of a RSP is not necessarily a RSP.

3.2 Robust Maximum Flow Problem (RMFP)

Based on [2], in this section we discuss the problem of finding a robust maximum flow. We first recall the maximum flow problem. Let $G = (\mathcal{V}, \mathcal{A})$ be a directed graph, let $s, t \in \mathcal{V}$ and let $c : \mathcal{A} \rightarrow \mathbf{Q}_+$ be a capacity function. The objective in the maximum flow problem is to find an $s - t$ flow of maximum value under c . Adding

an arc from t to s with $c_{ts} = \infty$, the maximum $s - t$ flow problem can be formulated as

$$\max\{x_{ts} : Ax = 0, 0 \leq x \leq c\}, \quad (13)$$

where A is the node-arc incidence matrix of size $|\mathcal{V}| \times |\mathcal{A}|$ and x is the vector of flow variables. In the case of uncertain maximum flow problem, the natural assumption is that the network $G = (\mathcal{V}, \mathcal{A})$ is fixed, as well as the nodes s and t . Thus, the uncertainty occurs only in the vector c of arc capacities. We assume $c \in \mathcal{U}$, where $\mathcal{U}$ is an uncertainty set for c . The robust counterpart of the RMFP can be stated as

$$\max\{x_{ts} : Ax = 0, 0 \leq x \leq c, \forall c \in \mathcal{U}\}.$$

Thus, the objective is to find the maximum value of a flow that satisfies $x \leq c$ for all $c \in \mathcal{U}$, where $c_{ts} = \infty$. Using the same types of uncertainty sets as presented in (9) and (11), we have the following results.

Theorem 3. *For the RMFP with box uncertainty as given by (9), the robust counterpart is simply*

$$\max\{x_{ts} : Ax = 0, 0 \leq x \leq l\}. \quad (14)$$

Theorem 4. *For the RMFP with ellipsoidal uncertainty as given by (11), the robust counterpart is*

$$\max\{x_{ts} : Ax = 0, 0 \leq x_a \leq c_a^n - \|\mathcal{Q}_a\|, a \in \mathcal{A}\},$$

where $\mathcal{Q}_a$ is the row of $\mathcal{Q}$ corresponding to arc a .

In both cases, the RMFP is a usual maximum flow problem with modified arc capacities, and hence the RMFP can be solved in polynomial time.

3.3 Robust Resistance Electrical Network Topology Design (RNTD) Problem

A resistance network is an electrical network comprised of two-terminal linear resistors only. We consider the situation where the in- and output currents at the nodes are given. Our aim is to design a resistance network that minimizes the dissipation, i.e., the energy consumption of the network. We call this the resistance network topology design (RNTD) problem. We present three different cases of the RNTD problem, namely the single-current case, the multi-current case and the robust RNTD problem (for details we refers to [4]).

Theorem 5. *For a given external current f , the following is the nonlinear problem for the single-current case.*

$$\text{Diss}_f(g) = \min_{g, y} \{f^T y : A(g)y = f, e^T g \leq \omega, g \geq 0\}, \quad (15)$$

where y is the vector of the potentials, g is the vector of conductance values and $A(g) = B \text{diag}(g) B^T$ is the conductance matrix, B is the structural matrix of the network.

Theorem 6. *Optimal solutions of the following linear problem and its dual can be used to construct an optimal solution to the nonlinear model (15).*

$$\max_y \{f^T y : |y^T b_{vw}| \leq 1, \forall (v, w) \in \mathcal{A}\}. \quad (16)$$

The dual problem of (16) is given by

$$\min_x \{ \|x\|_1 = \sum_{\{v,w\} \in \mathcal{A}} |x_{vw}| : \sum_{\{v,w\} \in \mathcal{A}} x_{vw} b_{vw} = f \}, \quad (17)$$

where b_{vw} denotes the column of B corresponding to arc (v, w) .

Using a simple variational principle (for details, see [4]), we obtain the semidefinite model for the RNTD problem.

Theorem 7. *Assume that the set $\mathcal{F}$ of current scenarios is a finite set $\mathcal{F} = \{f_1, \dots, f_k\}$. Then the following is the RNTD model for the multi-current case.*

$$\min_{\tau, g} \{ \tau : \begin{bmatrix} \tau & f_j^T \\ f_j & A(g) \end{bmatrix} \succeq 0, j = 1, \dots, k, e^T g \leq \omega, g \geq 0 \}. \quad (18)$$

The design variables are the conductance vector $g_i \in \mathbf{R}_+^n, i = 1, \dots, n$ (n is the number of arcs) and $\tau \in \mathbf{R}$.

Theorem 8. *Assume the set of external currents $\mathcal{F}$ as an ellipsoid:*

$$\mathcal{F} = \{f = Qz : z^T z \leq 1\}, \quad (19)$$

where Q is a matrix of size $m \times p$ and $z \in \mathbf{R}^p$ for some p . Then the robust RNTD problem is

$$\min_{\tau, g} \{ \tau : \begin{bmatrix} \tau I_p & Q^T \\ Q & A(g) \end{bmatrix} \succeq 0, e^T g \leq \omega, g \geq 0 \}. \quad (20)$$

This model finds the network which is best able to withstand *all* the currents in the ellipsoidal set of currents as given by (19).

4 Conclusions

We have discussed in this paper how to use conic optimization to model three robust design problems. These examples show that CO is indeed a powerful tool for the mathematical modelling of inherently nonlinear problems. Furthermore, CO enables us to model the robustness of a design in a computationally tractable way that opens the way to more new applications. Finding more new applications can be considered as an interesting challenge for the future.

Acknowledgement

Financial support was provided by the Royal Netherlands Academy of Arts and Sciences (KNAW) in the framework of the Scientific Programme Indonesia-Netherlands (SPIN).

References

1. A. Ben-Tal and A. Nemirovski. Robust optimization—methodology and applications. *Math. Program.*, 92(3, Ser. B):453–480, 2002.
2. D. Chaerani, C. Roos. The Robust Maximum Flow Problem and some examples, *Proceedings International Conference of Applied Mathematics 2005*, ITB, Indonesia, 2005.
3. D. Chaerani, C. Roos, A. Aman. The Robust Shortest Path Problem by means of Robust Linear Optimization., In H. Fleuren, D. den Hertog, and P. Kort editors, *Operations Research Proceedings 2004*, pages 335-342, Springer Verlag, Berlin, Heidelberg, New York, 2005.
4. C. Roos, Y. Bai, and D. Chaerani. Robust electrical network topology design by conic optimization. *Submitted to Optimization and Engineering*, 2005.
5. Y. Nesterov, A. Nemirovskii, Interior-point polynomial algorithms in convex programming, *SIAM Studies in Applied Mathematics*, Philadelphia, PA, 1994.

Polynomial Algorithms for Some Hard Problems of Finding Connected Spanning Subgraphs of Extreme Total Edge Weight

Alexey Baburin and Edward Gimadi

Sobolev Institute of Mathematics SB RAS, pr. Koptzyga 4, Novosibirsk, Russia
ababur@sicex.ru
gimadi@math.nsc.ru

Summary. Several hard optimization problems of finding spanning connected subgraphs with extreme total edge weight are considered. A number of results on constructing polynomial algorithms with performance guarantees for these problems is presented.¹

1 Introduction

This paper is devoted to address a subject of several previous works by the authors on problems of finding a restricted spanning subgraph of extreme total edge weight in a given graph. Restrictions imply connectivity of the resulting subgraph and predefined values of degrees of its vertices.

One of the most popular problems of the kind (when the degrees of all vertices equal two) is TSP [16]. The problem is MAX SNP-hard: existence of a polynomial approximation scheme for it yields $P = NP$. The Euclidean TSP in k -dimensional Euclidean space is also NP -hard. Although the status of the problem for $k = 2$ remains unclear. However there exist polynomial asymptotically optimal algorithms for solving the deterministic Maximum Euclidean TSP [2, 15]. We imply different restrictions on the resulting subgraph thus considering several generalizations of TSP.

We present polynomial approximation algorithms with performance guarantees for the problem of finding a connected spanning subgraph of maximum total edge weight with given (arbitrary) vertex degrees. Performance guarantees are proved for three cases of edge weights of the input graph: arbitrary non-negative weights, weights satisfying triangular inequality, Euclidean distances [10], [4]. The problem of finding a regular connected subgraph of extreme total edge weight in a complete graph when the weights of the edges are random variables is also considered. Polynomial asymptotically optimal algorithm is presented for maximization and minimizations variants of the problem [4].

¹ Research was supported by Russian Foundation for Basic Research (project 05-01-00395) and INTAS (grant 04-77-7173).

Another problem considered is a problem of finding several edge-disjoint Hamiltonian circuits of extreme total edge weight. The first mention of the problem called the m -Peripatetic Salesman Problem (m -PSP) in the literature came in [13]. Recently we developed several polynomial algorithms with performance guarantees for solving the maximum 2-PSP, and the metric minimum 2-PSP (cases of equal and independent weight functions of edges for measuring the weights of two Hamiltonian circuits constructed) [1], [3].

We also consider a maximization variant of Euclidean m -PSP and outline an approximation algorithm for solving it.

2 Finding a Spanning Connected Subgraph of Extreme Total Edge Weight and Given Vertex Degrees

Let $G(V, E)$ be a complete n -vertex undirected graph without loops with a non-negative weight function w of edges. There are known integers d_i ($i = 1, \dots, n$), $1 \leq d_i < n$.

In [12] the problem of finding a realization of a set of integers as degrees of the vertices in a subgraph G' of G is formulated. (This set is called a graphical partition of a number p , where $p = \sum_{i=1}^n d_i$). It is clear that for every such realization the number p is even and $d_i < n$ for each $i = 1, \dots, n$. These conditions are not sufficient. A constructive criterion of realizability of a set of integers was developed in [11].

An optimization appearance of the problem was described in [8] by means of reduction to the problem of finding a maximum-weight matching.

The problem of finding a maximum-weight spanning connected subgraph with given vertex degrees appeared in [10]. An approximation algorithm was presented for the metric case of the problem. The algorithm only worked for the case when all integers d_i were even. In [4] a new approximation algorithm was developed. It did not require d being even and worked for general case of the problem.

Theorem 1. [4] *An approximate solution for the problem can be found with running-time $O(Mn^2)$, where M is the number of edges in the subgraph being built. The performance ratio of the solution does not exceed $2/(d^2 + d)$, where $d = \min\{d_i | i = 1, \dots, n\}$. In the metric case of the problem the performance ratio does not exceed $1/(d^2 + d)$.*

3 Finding d -regular Connected Subgraph in a Complete Graph with Random Edge Weights

Let $G(V, E)$ be a complete n -vertex undirected graph without loops with a weight function $w : E \rightarrow R^+$ of edges and d , $1 < d < n$, be a given number. The problem is to find a d -regular connected subgraph $\tilde{G}$ of G of maximal (or minimal) total edge weight. A solution always exists for the case of even nd and $d < n$. The case of $d = 2$ is the Maximum Traveling Salesman Problem. A survey of results on the later problem can be found in [16].

Suppose an approximation algorithm A solves an optimization problem of maximization. By $F_A(I)$, $F^*(I)$ we denote the values of the objective function when solving an individual problem I by algorithm A and precisely respectively. Algorithm A has a *relative error* ε_n and a *fault probability* δ_n in the class K_n of n -dimensional problems if for all n the following probabilistic inequality holds:

$$\Pr\left\{\left|\frac{F^*(I) - F_A(I)}{F^*(I)}\right| > \varepsilon_n\right\} \leq \delta_n.$$

An algorithm is called *asymptotically optimal* in the problem class $K = \cup\{K_n | n = 1, 2, \dots\}$ if there exist its performance estimates ε_n, δ_n such that $\varepsilon_n \rightarrow 0, \delta_n \rightarrow 0$ as $n \rightarrow \infty$.

Consider the class $\tilde{K}$ of weight matrixes (random instances of the problem) with elements independently distributed in a segment $(0,1)$ with the same distribution function of *minorizing* type: $F(x) \leq x$ for $0 \leq x \leq 1$.

Theorem 2. [4] *If $d = o(n)$, then an asymptotically optimal solution for the problem of finding d -regular connected subgraph of maximal total edge weight on class $\tilde{K}$ can be found with running-time $O(n^2)$.*

Similarly we can define distribution functions of *majorizing* type on a segment (a_n, b_n) , $a_n > 0$. In this case the same results can be obtained for a minimization problem if another restrictions on the input is implied:

$$\frac{b_n}{a_n} = o\left(\frac{n/d}{\log(n/d)}\right).$$

4 Finding two Edge-Disjoint Hamiltonian Circuits

Let $G = (V, E)$ be a complete undirected graph with n vertices. The edges of the graph are weighted with two functions $w_1 : E \rightarrow R$, $w_2 : E \rightarrow R$. By $W_i(H)$ we denote $W_i(H) = \sum_{e \in H} w_i(e)$, where $i = 1, 2$.

We consider a problem of finding two edge-disjoint Hamiltonian circuits $H_1 \subset E$ and $H_2 \subset E$ such that their total weight $W_1(H_1) + W_2(H_2)$ is minimal (or maximal). We will refer the problem as *the Single 2-PSP*, if $w_1 \equiv w_2$, and *the Dooble 2-PSP*, if w_1 and w_2 are independent weight functions.

In [7] it is shown that the Single 2-PSP is *NP*-complete. This result implies that Single 2-PSP is *NP*-hard both in maximization and minimization variants. In [3] *NP*-hardness of the Dooble 2-PSP was established also.

4.1 Single 2-PSP_{max}

A polynomial approximation algorithm with performance ratio of $3/4$ for solving the Single 2-PSP_{max} was presented in [1]. First of all a cubic (if n is even) of "almost" cubic (if n is odd) subgraph K of G of maximal summary edge weight is constructed. Then K is split into partial tour and 2-matching. The subgraphs then are modified into two partial tours by regrouping their edges. The resulting partial tours can be fulfilled into two edge-disjoint Hamiltonian circuits by adding other edges. The

weight of the solution is at least $(3/4)W^*$, where W^* is the weight of the optimal solution.

The performance guarantee relies on the facts that the total edge weight of K is at least $(3/4)W^*$ and all the edges of K were included into the solution. The running time of the algorithm is determined by the complexity of finding K in G and is bounded by $O(n^3)$ as described in [9].

4.2 Metric 2-PSP_{min}

It is supposed that the triangle inequality holds. In [3] $O(n^3)$ -algorithms was presented with performance ratios $9/4$ for the Single Metric 2-PSP_{min}, and $12/5$ for the Dooble Metric 2-PSP_{min}. Later the same performance ratio for the Single Metric 2-PSP_{min} was also announced in [6].

We will call by the 2-PSP_{min}(1,2) the particular case of the problem when the weights of individual edges take the values 1 or 2. Evidently in this case we have also the metric problem.

4.3 The Single 2-PSP_{min}(1,2)

In [6] performance ratio of about $197/144$ was announced for this problem in assumption that performance ratio $7/6$ holds for solution of TSP_{min}(1,2) on input graph, found by algorithm presented in [14]. The following statement gives the better performance ratio.

Theorem 3. *Let ρ_1 be the performance ratio of an algorithm A solving TSP_{max}(1,2). Then the Single 2-PSP_{min}(1,2) can be solved with a performance ratio at most $(1 + \rho_1)$ in time determined by the algorithm A .*

Thus an approximate solution with total weight of at most $5/4$ times the optimal for the Single 2-PSP_{min}(1,2) can be found in $O(n^3)$ running-time.

4.4 The Double 2-PSP_{min}(1,2).

Theorem 4. *Let ρ_2 be the performance ratio of an algorithm A solving TSP_{min}(1,2). Then the Dooble 2-PSP_{min}(1,2) can be solved with a performance ratio at most $(1 + \rho_2/2)$ in time determined by the algorithm A .*

Corollary *It is possible to find a feasible solution of the Dooble 2-PSP_{min}(1,2) with performance ratio of at most $11/7$ in $O(n^3)$ -time.*

The statement relies on the ratio $\rho_2 = 8/7$ for the approximate solution for TSP_{min}(1,2), found by algorithm from [5]. Though the running time of the algorithm used is polynomial, it is very high: $O(n^{K+4})$, where the constant K in [5] is equal to 21. Thus, using an algorithm from [14] for TSP_{min}(1,2) (with the ratio $7/6$) implies slightly greater value of $19/12$, however the running-time of the algorithm is at most $O(n^3)$, i.e. much smaller.

5 Euclidean m -PSP $_{max}$ in Space $\mathbf{R}^k$

At first we give an outline of asymptotically optimal algorithm for solving Euclidean 1-PSP $_{max}$ (i.e Euclidean TSP $_{max}$) [2, 15, 16].

Lemma 1. [15]. *Let S be an arbitrary set of t intervals from $\mathbf{R}^k$. Then the minimal angle between any pair of intervals from the set is not greater than α_{kt} , such that $\sin^2(\alpha_{kt}/2) \leq \gamma_k t^{-2/(k-1)}$, where γ_k depends only on the dimension of $\mathbf{R}^k$. So $\alpha_{kt} \rightarrow 0$ as $t \rightarrow \infty$.*

Lemma 2. [15]. *Let $\alpha \leq \pi/2$ be an angle between a pair of intervals $I_j = (x_j, y_j)$, $I_l = (x_l, y_l)$ from $\mathbf{R}^k$. Then $\max\{w(I_j), w(I_l), \cos \frac{\alpha}{2}(w(I_j) + w(I_l))\} \leq \max\{w(x_j, x_l) + w(y_j, y_l), w(x_j, y_l) + w(y_j, x_l)\}$.*

Algorithm for the Euclidean TSP $_{max}$

1. Split edges (intervals) of maximum weight matching M into two subsets of $(t-1)$ "light" and the rest "heavy" edges, $2 \leq t \leq n/4$. An edge or a series of edges, that form a cycle in the graph, we refer as a *collection*. *Leading* and *driven* edges are chosen in each collection.

2. Repeat the following procedure $\mathcal{P}$ with one-element collections consisting of heavy edges $\mathcal{P}'$ while the number of collections becomes of t .

Procedure $\mathcal{P}$:

P_1 : Pick out 2 collections with a minimal angle between their leading edges;
 P_2 : Patch the chosen collections by a pair of edges that connect them with maximum summary weight. One of the driven edges we denote as leading.

3. The set of t collections obtained using the procedure $\mathcal{P}$ alternate with $t-1$ light edges. The sequence of thus formed collections we are patched into a Hamiltonian cycle using the application of $t-1$ operations P_2 .

Theorem 5. *For $t^* = \lceil n^{(k-1)/(k+1)} \rceil$ Euclidean TSP $_{max}$ can be solved in $O(n^{5/2})$ -time asymptotically optimal with the relative error $\varepsilon_n = O(n^{-2/(k+2)})$.*

Euclidean m -PSP $_{max}$ can be solved with good approximation in polynomial time, repeating the algorithm described above on steps $i = 1, \dots, m$ with the same matching M .

To avoid reuseage of the edges, restrictions of not using $2n(i-1)$ pairs of matching edges connected by one of the edge from the previous Hamiltonian cycles $H_1, \dots, H_{i-1}$ are implied while constructing the next circuit H_i .

References

1. Ageev A., Baburin A, Gimadi E. (2006) Polynomial algorithm with performance guarantee of $3/4$ for a problem of finding two edge-disjoint Hamiltonian circuits of maximal total edge weight (in russian). Discrete analysis and oper. research. Ser. 1., V. 13, 2:10–20.
2. Baburin E., Gimadi E. (2002) On asymptotical optimality of algorithm for solving maximum weight Euclidean TSP (in russian). Discrete analysis and operations research. Ser. 1., V. 9, 4: 23–32.

3. Baburin A., Gimadi E., Korkishko N. (2004) Approximation algorithms for finding two edge-disjoint Hamiltonian circuits of minimal total edge weight (in russian). *Discrete analysis and oper. research. Ser. 2.*, V. 11, 1:11–25.
4. Baburin E., Gimadi E. (2005) Approximation algorithms for finding a maximum-weight spanning connected subgraph with given vertex degrees. *Oper. Res. Proc. 2004, Intern. Conf. OR 2004*, Tilburg. Springer, Berlin, 343–351.
5. Berman P., Karpinski M. (2006) $8/7$ -approximation algorithm for (1,2)-TSP. *Proc. 17th ACM-SIAM SODA*, 641–648.
6. Croce F.D., Pashos V.Th., Calvo R.W. (2005) Approximating the 2-peripatetic salesman problem. 7th Workshop on Modelling and Algorithms for Planning and Scheduling Problems MAPS 2005. (Siena, Italy, June 6-10). 114–116.
7. De Kort J.B. J.M. (1991) Lower bounds for symmetric K -peripatetic salesman problems. *Optimization*. V. 22, 1: 113–122.
8. Edmonds J., Johnson E. L. (1970) Matchings: a well solvable class of integer linear programs. *Combinatorial Structures and Their Applications*. New York: Gordon and Breach. 89–92.
9. Gabow H. N. (1983) An efficient reduction technique for degree-constrained subgraph and bidirected network flow problems. *Proc. 15th annual ACM symposium on theory of computing*, (Boston, April 25-27, 1983). New York: ACM, 448–456.
10. Gimadi E.Kh., Serdyukov A.I. (2001) Problem of Finding the Maximal Spanning Connected Subgraph with Given Vertex Degrees. *Oper. Res. Proceed. 2000* (Eds Fleishman B. at al.) Springer, Berlin, 55–59.
11. Havel V. (1955) A note to question of existance of finite graphs, *Casopis Pest Mat.*, V. 80, 477–480.
12. Harary F. (1969) *Graph Theory*, Addison-Wesley, Reading, Massachusetts.
13. Krarup J. (1975) The peripatetic salesman and some related unsolved problems. *Combinatorial programming: methods and applications* (Proc. NATO Advanced Study Inst., Versailles, 1974). 173–178.
14. Papadimitriou C. H., Yannakakis M. (1993) The traveling salesman problem with distance One and Two. *Math. Oper. Res.* V. 18, 1:1–11.
15. Serdyukov A.I. (1987) An Asymptotically Optimal Algorithm for Solving Max Euclidean TSP. *Metodi Celochislennoi Optimizacii (Upravliaemie Sistemi)*, Novosibirsk. 27:79–87.
16. *The Traveling Salesman Problem and its Variations*. Gutin G., Punnen A.P. (eds). Kluwer Academic Publishers, Dordrecht / Boston / London. 2002.

**Econometrics, Game Theory and Mathematical
Economics**

A Multidimensional Poverty Index

Gerhard Kocklaeuner

FB Wirtschaft, FH Kiel, Germany
gerhard.kocklaeuner@fh-kiel.de

1 Introduction

In multidimensional poverty measurement $k \geq 2$ different quantitative basic needs variables $y_j, j = 1 \dots k$ have to be considered. Let the respective variables be substitutable. Then for a single person $i = 1 \dots n$ poverty can be aggregated across variables as follows:

$$C_{1i}^* = \left(\frac{1}{k} \sum_{j=1}^k p_{ij}^{*\alpha} \right)^{\frac{1}{\alpha}}, \alpha \geq 2 \quad (1)$$

(see Kocklaeuner (2002) with respect to an unidimensional ethical poverty index aggregating poverty across persons).

In equation (1) $p_{ij}^* = \max[p_{ij}, 0]$, i.e. a censored poverty gap ratio $p_{ij} = \frac{z_j - y_{ij}}{z_j}$, where z_j denotes a finite positive poverty line corresponding to variable y_j and y_{ij} denotes a finite non-negative observation of this variable. Note that C_{1i}^* is positive if y_{ij} is less than z_j for some j (weak definition of the poor).

In the following the intrapersonal perspective of the index C_{1i}^* will be extended to an interpersonal one. The poverty index resulting will aggregate poverty both across persons and across variables. It will be shown to satisfy essential poverty axioms.

2 The Poverty Index C_1^*

In Equation (1) an α -mean is used to measure poverty for a single person. When it comes to combine such measures for a group of n persons the application of a corresponding β -mean seems to be natural. This leads to the multidimensional poverty index

$$C_1^* = \left(\frac{1}{n} \sum_{i=1}^n C_{1i}^{*\beta} \right)^{\frac{1}{\beta}}, \beta \geq 2. \quad (2)$$

Note that a poverty index like $C_1^{*\beta}$ has been proposed by Bourguignon and Chakravarty (1998a and 1998b). The special case $\alpha = \beta$ has been used by Foster

et al. (2005) in the context of measuring human development. The human poverty index HPI-1 of UNDP (see e.g. UNDP (2001:273)) arises from C_1^* by taking a representative person, i.e. $n = 1$, and $\alpha = k = 3$ in equation (2).

Contrary to poverty indexes in the line of Foster et al. (1984) the index C_1^* proves to be an ethical poverty measure. It depends on social welfare functions. With respect to equation (1) such a function gives reasons for the special measure of inequality

$$B_{pi}^{*\alpha} = \frac{1}{\bar{p}_i} \left(\frac{1}{k} \sum_{j=1}^k p_{ij}^{*\alpha} \right)^{\frac{1}{\alpha}} - 1, \alpha \geq 2 \quad (3)$$

(compare Kocklaeuner (2002), equation (3)). The application of equation (3) to the β -mean of equation (2) is straightforward.

Obviously, the index C_1^* can be rewritten as

$$C_1^* = \frac{1}{n^{\frac{1}{\beta}}} \left(\sum_{i=1}^n (C_{1i}^* - 0)^{\beta} \right)^{\frac{1}{\beta}}, \beta \geq 2. \quad (4)$$

Equation (4) shows the index C_1^* to be a normalized Minkowski distance between the existing poverty- and an ideal no-poverty-case (see Subramanian (2004)). Again, an adaptation of equation (4), now with respect to the index C_{1i}^* , is straightforward.

The index C_1^* satisfies classical poverty axioms known from unidimensional poverty measurement (see e.g. Zheng (1997)). Besides normalization to the interval $[0,1]$ these are the axioms of strong continuity, focus, symmetry, replication invariance, scale invariance, increasing poverty lines, non-poverty growth, strong monotonicity, strong transfer, strong transfer sensitivity, increasing poverty aversion, and subgroup consistency. All these axioms hold just because the index C_1^* is defined in a two-step-procedure where a single step refers to a single type of aggregation.

Strong continuity combines continuity of C_1^* and of $\frac{\partial C_1^*}{\partial y_{ij}}$ with respect to y_{ij} especially at $y_{ij} = z_j$. Scale invariance implies unit consistency (see Zheng (2005)). Strong transfer sensitivity combines strong continuity and weak transfer sensitivity (see Zheng (1999)). Here the condition of $\frac{\partial^3 C_1^*}{\partial y_{ij}^3}$ being negative for y_{ij} less than z_j leads to α greater than 2 and β greater than 2α . In case of $z_j = z, j = 1 \dots k$ corresponding transfers are absolute intrapersonal transfers from variable y_j to another variable $y_{j'}$ or from variable $y_{j'}$ to another variable $y_{j''}$. In case of different poverty lines the respective transfers have to be relative to z_j . This qualification also refers to increasing poverty aversion where in equation (1) α can be interpreted as a parameter of poverty aversion (see Seidl (1988)).

Contrary to poverty indexes following Foster et al. (1984) the index C_1^* is not decomposable. Similarly the indexes $C_{1i}^{*\alpha}$ and $C_1^{*\beta}$ do not satisfy increasing poverty aversion. However, subgroup consistency of the index C_1^* with respect to intrapersonal groups of variables depends on decomposability of $C_{1i}^{*\alpha}$.

On the interpersonal level a mean-preserving transfer can include separate transfers for different variables. It can be formalized by multiplying the (n,k) -matrix $Y = (y_{ij})$ of observations by a bistochastic (n,n) -matrix B whereby permutation matrices are excluded and poverty lines have to remain constant. Consequently, such a transfer while decreasing interpersonal differences can increase the number of poor persons. In case of β greater than α the first type of change will dominate. So the index C_1^* still satisfies the strong transfer axiom if $\beta \geq \alpha$. I.e. the value of C_1^* will

decrease in the course of the respective transfer. Interestingly this result is contrary to Tsui (2002) who pretends that "there does not exist any subgroup consistent multidimensional poverty index satisfying PDMT" (the poverty decreasing minimal transfer axiom). However, in his evidence Tsui ignores the possibility of censoring being essential for C_1^* .

A special type of mean-preserving interpersonal transfer is a poverty-non-decreasing rearrangement. Such a transfer only concerns row i and row i' (i less than i') of matrix Y which are assumed not to be identical. After the transfer row i contains $\min[y_{ij}, y_{i'j}]$ and row i' contains $\max[y_{ij}, y_{i'j}]$, $j = 1 \dots k$ as elements. The value of C_1^* now increases as a result of such a transfer if β is greater than α . Thus the index C_1^* satisfies the axiom of poverty-increasing rearrangements. Evidence of this can be found when showing the index $C_1^{*\beta}$ to be an L-superadditive function of observations y_{ij} and $y_{i'j}$ ($j' \neq j$) being less than z_j and $z_{j'}$ respectively. L-superadditivity proves to be equivalent to a positive second mixed partial derivative (see Tsui (2002:79)). Here $\frac{\partial^2 C_1^{*\beta}}{\partial y_{ij} \partial y_{i'j'}}$ is positive if β is greater than α .

References

1. UNDP (2001) Bericht ueber die menschliche Entwicklung. UNO-Verlag, Bonn
2. Seidl CH (1988) Poverty Measurement - A Survey. In: Boes D et al. (eds) Welfare and Efficiency in Public Economics:71-147. Springer, Berlin Heidelberg New York
3. Foster JE et al. (1984) A Class of Decomposable Poverty Measures. *Econometrica* 52:761-766
4. Foster JE et al. (2005) Measuring the Distribution of Human Development: methodology and an application to Mexico. *Journal of Human Development* 6:5-29
5. Kocklaeuner G (2002) Revisiting Two Poverty Indexes. *Allgemeines Statistisches Archiv* 86:299-305
6. Tsui K (2002) Multidimensional Poverty Indices. *Social Choice and Welfare* 19:69-93
7. Zheng B (1997) Aggregate Poverty Measures. *Journal of Economic Surveys* 11:123-162
8. Zheng B (1999) On the Power of Poverty Orderings. *Social Choice and Welfare* 16:349-371
9. Bourguignon F, Chakravarty SR (1998a) A Family of Multidimensional Poverty Measures. DELTA document No.98-03
10. Bourguignon F, Chakravarty SR (1998b) The Measurement of Multidimensional Poverty. DELTA document No.98-12
11. Subramanian S (2004) A Re-scaled Version of the Foster-Greer-Thorbecke Poverty Indices based on an Association with the Minkowski Distance Function. UNU-WIDER Research Paper No.2004/10
12. Zheng B (2005) Unit-consistent Poverty Indices. Forthcoming

Parameter Estimation for Stock Models with Non-Constant Volatility Using Markov Chain Monte Carlo Methods

Markus Hahn, Wolfgang Putschögl, and Jörn Sass

RICAM, Austrian Academy of Sciences, Linz, Austria*

`markus.hahn@oeaw.ac.at`

`wolfgang.putschoegl@oeaw.ac.at`

`joern.sass@oeaw.ac.at`

Summary. We consider a model for a financial market where the asset prices satisfy a stochastic differential equation. For the volatility no new source of randomness is introduced, but the volatility at each time depends deterministically on all previous price fluctuations. Such non-constant volatility models preserve the completeness of the market while they allow for many attractive features.

We deal with the estimation of the occurring parameters given discretely observed data deriving a suitable Markov chain Monte Carlo framework. We illustrate this with an example dealt with in [2] where the drift term is a continuous time Markov chain with finitely many states (which cannot be observed directly) and the volatility is given by the Hobson-Rogers model [3]. We also present numerical results.

1 Introduction

For a finite time horizon $T > 0$ we look at a stock market model. The stock returns are given by $dR_t = \mu_t dt + \sigma_t dW_t$, where the drift process $\mu = (\mu_t)_{t \in [0, T]}$ is independent of the Brownian motion W . For the volatility $\sigma = (\sigma_t)_{t \in [0, T]}$ we consider so-called non-constant processes that are functions of observable factor processes, i.e. they depend deterministically on the past stock prices, and hence introduce no new source of randomness. Thus we have a complete market which is not true for (fully) stochastic volatility models where a non-traded source of risk is included. We describe a very general and flexible framework to calibrate such models to market data using Markov chain Monte Carlo (MCMC) methods popular in finance since [4].

One example used as an illustration is the Hobson-Rogers (HR) model introduced in [3]. There the volatility is a deterministic function of the past underlying prices represented by finitely many offset functions. This model accounts for the possibility of volatility smiles and skews as well as term structures, even using a model based

* supported by FWF Project P17947-N12

on the first offset function only. In [5] it is found that the fit of the HR model is generally better than that of the Heston model.

2 A Stock Model with Non-Constant Volatility

Let $(\Omega, \mathcal{A}, P)$ be a complete probability space, $T > 0$ the terminal trading time, and $\mathcal{F} = (\mathcal{F}_t)_{t \in [0, T]}$ a filtration in $\mathcal{A}$ satisfying the usual conditions, i.e. $\mathcal{F}$ is right-continuous and contains all P -null sets. We consider a riskless asset (bond) with non-negative interest rates $(r_t)_{t \in [0, T]}$ and n risky assets (stocks). The dynamics of the n -dimensional price process $S = (S_t)_{t \in [0, T]}$ of the stocks is given by

$$dS_t = \text{Diag}(S_t)(\mu_t dt + \sigma_t dW_t), \quad S_0^i > 0, \quad (1)$$

where $W = (W_t)_{t \in [0, T]}$ is an n -dimensional Brownian motion w.r.t. $\mathcal{F}$. The drift process $\mu = (\mu_t)_{t \in [0, T]} \in \mathbf{R}^n$ is independent of W , the volatilities $(\sigma_t)_{t \in [0, T]} \in \mathbf{R}^{n \times n}$ are non-singular, and both μ and σ are progressively measurable w.r.t. $\mathcal{F}$. The return process $R = (R_t)_{t \in [0, T]}$ associated with the stock is defined by $dR_t = \text{Diag}(S_t)^{-1} dS_t$, the excess return process is defined by $d\tilde{R}_t = dR_t - r_t dt$.

Without loss of generality we can assume σ_t to be a lower triangular matrix with positive diagonal. Since $\sigma_t \sigma_t^\top dt = d[R]_t$ and σ_t is given uniquely as the square root, $(\sigma_t)_{t \in [0, T]}$ is adapted to $\mathcal{F}^S$ as well as to $\mathcal{F}^R$. Here $[\cdot]$ denotes the quadratic variation, and $\mathcal{F}^X = (\mathcal{F}_t^X)_{t \in [0, T]}$ denotes the filtration of augmented σ -algebras generated by the $\mathcal{F}$ -adapted process $X = (X_t)_{t \in [0, T]}$.

Next, we introduce the risk neutral probability measure $\tilde{P}$. The density process is defined by $dZ_t = -\sigma_t^{-1}(\mu_t - r_t)Z_t dW_t$, $Z_0 = 1$. We assume that Z is a martingale and define $\tilde{P}$ by $d\tilde{P} = Z_T dP$. Girsanov's theorem guarantees that $\tilde{W}_t = W_t + \int_0^t \sigma_s^{-1}(\mu_s - r_s) ds$ is a $\tilde{P}$ -Brownian motion. The excess return process is a martingale under $\tilde{P}$ satisfying $d\tilde{R}_t = \sigma_t d\tilde{W}_t$.

We model the dynamics of the volatilities under $\tilde{P}$ using an $\mathcal{F}^S$ -adapted process ξ which leads to a Markovian structure introducing the m -dimensional factor process $\xi = (\xi_t)_{t \in [0, T]}$ with dynamics

$$d\xi_t = \nu(\xi_t)dt + \tau(\xi_t)d\tilde{W}_t, \quad (2)$$

where ν and τ are $\mathbf{R}^m$ and $\mathbf{R}^{m \times n}$ -valued. The volatility process takes the form $\sigma_t = f_\sigma(\xi_t)$. We demand that ν , τ , and f_σ are measurable.

Remark 1. In (2) no new Brownian motion is introduced for the volatility model. Under standard Lipschitz and linear growth conditions on ν , τ , f_σ the system consisting of $d\tilde{R}_t = \sigma_t d\tilde{W}_t$ and (2) has a strong solution $(\tilde{R}, \xi)$ (which is Markovian) and the model is complete w.r.t. $\mathcal{F}_T^S$ (cf. [2]).

3 A General MCMC Framework

We construct a block-wise Metropolis-Hastings (MH) sampler for the time-discretised model. To this end we have to augment the unknowns such that the complete-data likelihood (CDLH) is computable, fix prior distributions, and construct update schemes. Here we employ the original measure P .

3.1 Discretization

We use a discrete time approximation assuming that drift and volatility are constant between equidistant observation times $\kappa\Delta t$, $\kappa = 0, \dots, N$, $\Delta t > 0$, $N\Delta t = T$. We denote the discrete processes with dots, e.g. $\dot{\mu} = (\dot{\mu}_\kappa)_{\kappa=0, \dots, N-1}$, $\dot{\mu}_\kappa = \mu_{\kappa\Delta t}$, hence $\dot{\sigma}_\kappa = f_\sigma(\dot{\xi}_\kappa)$ and

$$\Delta\dot{R}_{\kappa+1} = \dot{\mu}_\kappa\Delta t + \dot{\sigma}_\kappa\Delta\dot{W}_{\kappa+1}, \quad \Delta\dot{\xi}_{\kappa+1} = \nu(\dot{\xi}_\kappa)\Delta t + \tau(\dot{\xi}_\kappa)\dot{\sigma}_\kappa^{-1}\Delta\dot{R}_{\kappa+1}. \quad (3)$$

3.2 Metropolis-Hastings Sampler

If the drift process μ is random and cannot be observed directly, the variables may need to be augmented with the hidden process determining μ . This is very useful e.g. for the latent state process in a hidden Markov model (HMM) like in Section 4 while it is not necessary e.g. for an Ornstein-Uhlenbeck process.

We partition the unknowns into two blocks M and Σ containing all parameters (and augmented variables) belonging to μ and σ , respectively. With Θ we denote the collection of all unknowns, i.e. the union of M and Σ .

Due to the Markovian nature of (R_t, ξ_t) the CDLH equals $P(\dot{R}|\Theta) = \prod_{\kappa=0}^{N-1} \varphi(\Delta\dot{R}_{\kappa+1}; \dot{\mu}_\kappa\Delta t, \dot{\sigma}_\kappa\dot{\sigma}_\kappa^\top\Delta t)$, where $\varphi(\cdot; \mu, \sigma\sigma^\top)$ denotes the multivariate normal density with mean μ and covariance $\sigma\sigma^\top$. Hence we can compute the posterior distribution $P(\Theta|\dot{R}) \propto P(\Theta)P(\dot{R}|\Theta)$ up to a factor.

We assume that M and Σ are independent a priori and fix appropriate prior distributions $P(M)$, $P(\Sigma)$.

Finally suitable proposal distributions $P(M, M')$, $P(\Sigma, \Sigma')$ have to be defined. Then the acceptance probability for M is given by

$\alpha_M = \min\{1, P(\dot{R}|M', \Sigma)P(M')P(M', M)/(P(\dot{R}|M, \Sigma)P(M)P(M, M'))\}$ and similarly for Σ . It may be useful to proceed stepwise on partitions of M and Σ . Note that from (3) we can compute the factor process and hence the volatility, so all occurring expressions can be evaluated.

4 An Example: HR Volatility and HMM Drift

In this section we present an application of the general methods described in the previous section specifying processes for the volatility and for the drift.

We consider a market consisting of one stock and a bond, where for notational simplicity we set $r_t = 0$, i.e. $R_t = \tilde{R}_t$. In [3] a class of stock price models is introduced where the instantaneous volatility is modeled in terms of exponentially weighted moments of the discounted historic log-prices. As a consequence present shocks in stock prices result in higher future volatility. The HR model is a non-constant volatility model leading to a complete market. For some empirical analysis of the model see [5] and the references therein. In the following we look at the model based on the first offset function (HR1 model). For the drift we use a continuous time HMM.

4.1 Hobson-Rogers First Order Volatility

As the factor process introduced in (2) we use the first offset function ξ ,

$$\xi_t = \int_0^\infty \lambda e^{-\lambda u} (\tilde{R}_t - \tilde{R}_{t-u}) du, \quad t \in [0, T], \quad (4)$$

satisfying the SDE $d\xi_t = d\tilde{R}_t - \lambda \xi_t dt$. Here $\lambda > 0$ is a parameter that determines the rate at which past information is discounted. The offset function at any time t depends on all the past values of the price process. Thus for the computation of the offset function ξ_t we need the path of $\tilde{R}$ on its past, where an appropriate initial value ξ_0 also has to be estimated. As proposed in [3] the volatility depending on the offset is defined as $\sigma_t = f_\sigma(\xi_t) = \underline{\sigma}(1 + \epsilon \xi_t^2)^{1/2}$. The minimum volatility is given by $\underline{\sigma} > 0$ and the parameter $\epsilon > 0$ scales the influence of the offset function. In contrast to [3] in the definition of the offset we use the excess return instead of the discounted log-prices and do not cut off at a maximal level of the volatility.

4.2 A Hidden Markov Model for the Drift

For the drift process of the returns we consider a HMM, i.e. μ is a stationary continuous time irreducible Markov chain, independent of W , and adapted to $\mathcal{F}$ with d states $b^1, \dots, b^d \in \mathbf{R}$. We represent μ as $\mu_t = b^\top Y_t$, where $Y = (Y_t)_{t \in [0, T]}$ is a stationary continuous time irreducible Markov chain, independent of W , and with state space $\{e_1, \dots, e_d\}$, the standard unit vectors in $\mathbf{R}^d$. The state process Y is further characterized by its rate matrix $Q \in \mathbf{R}^{d \times d}$, where $Q_{kl} = \lim_{t \rightarrow 0} \frac{1}{t} P(Y_t = e_l | Y_0 = e_k)$, $k \neq l$, is the jump rate or transition rate from e_k to e_l . Moreover $\lambda_k = -Q_{kk} = \sum_{l=1, l \neq k}^d Q_{kl}$ is the rate of leaving e_k , i.e. the waiting time for the next jump is exponentially distributed with parameter λ_k . The probability that the chain jumps to e_l when leaving e_k equals Q_{kl}/λ_k .

4.3 Calibration of the Model

We apply the methods from Section 3 to estimate parameters b , Q , ϵ , ξ_0 , $\underline{\sigma}$. First we discretize the continuous-time model, and second we describe the MH sampler based on data augmentation with the latent state process Y .

4.3.1 The Model in Discrete Time

Adopting the notation of Section 3, we first consider the discretised state process $\dot{Y} = (\dot{Y}_\kappa)_{\kappa=0, \dots, N-1}$. To allow jumps at the observation times $\kappa \Delta t$, $\kappa = 0, \dots, N$ only, we replace the rate matrix Q with the transition matrix $X \in \mathbf{R}^{d \times d}$, where $X_{kl} = P(Y_{t+\Delta t} = e_l | Y_t = e_k)$, i.e. $X = \exp(Q \Delta t)$. For the discrete time scheme we have $\dot{\sigma}_\kappa = \underline{\sigma}(1 + \epsilon \dot{\xi}_\kappa^2)^{1/2}$ and

$$\Delta \dot{R}_{\kappa+1} = \dot{\mu}_\kappa \Delta t + \dot{\sigma}_\kappa \Delta \dot{W}_{\kappa+1}, \quad \dot{\xi}_{\kappa+1} = \dot{\xi}_\kappa + \Delta \dot{R}_{\kappa+1} - \lambda \dot{\xi}_\kappa \Delta t. \quad (5)$$

Remark 2. From (5) we get $\dot{\xi}_\kappa = \Delta \dot{R}_\kappa$ for $\lambda \approx \Delta t^{-1}$, i.e. for large λ the offset process looks (up to a factor) very much like the return. Hence overestimating λ results in “over-fitting” the data. This makes the estimation of λ extremely difficult and it seems more reasonable to fix λ a priori which we prefer to putting a very restricting prior distribution on λ .

4.3.2 Metropolis-Hastings Sampler

We augment the unknowns with the hidden state process $\dot{Y}$ and get $\Theta = (b, X, \dot{Y}, \xi_0, \underline{\sigma}, \epsilon)$ which is blocked into $M = (b, X, \dot{Y})$ and $\Sigma = (\xi_0, \underline{\sigma}, \epsilon)$. For the updates we further block M into $b, X, \dot{Y}$.

For the CDLH we have $P(\dot{R}|\Theta) = \prod_{\kappa=0}^{N-1} \varphi(\Delta \dot{R}_{\kappa+1}; b^\top \dot{Y}_\kappa \Delta t, \dot{\sigma}_\kappa^2 \Delta t)$.

For the prior distribution we assume that $b, X, \dot{Y}_0, \xi_0, \underline{\sigma}, \epsilon$ are independent a priori, b^k and ξ_0 are normally distributed, $\underline{\sigma}$ and ϵ are Gamma distributed, Y_0 is uniformly distributed on $\{1, \dots, d\}$, and the rows X_k of the transition matrix are independent and Dirichlet distributed with parameters $(g_{k1}, \dots, g_{kd})$. As $\underline{\sigma}, \epsilon, \lambda > 0$, the Gamma distribution should be an appropriate prior distribution. The choice of the Dirichlet prior for the rows of the transition matrix allows to perform a Gibbs step for the update of X .

For the update of b we perform a Gibbs step drawing each component separately as $b^{k'} \sim N(\bar{m}^k, \bar{s}_k^2)$, where $\bar{s}_k^{-2} = s_0^{-2} + \sum_{\kappa=0}^{N-1} \mathbb{1}_{\{\dot{Y}_\kappa=k\}} \dot{\sigma}_\kappa^{-2}$ and $\bar{m}^k = (m_0^k s_0^{-2} + \sum_{\kappa=0}^{N-1} \mathbb{1}_{\{\dot{Y}_\kappa=k\}} \frac{\Delta \dot{R}_{\kappa+1}}{\Delta t} \dot{\sigma}_\kappa^{-2}) \bar{s}_k^2$ for $b^k \sim N(m_0^k, s_0^2)$ a priori.

Given $\dot{Y}$, the transition matrix depends on the number of jumps J_{kl} from state k to l only. As $P(J_k | X_k)$ is a multinomial distribution with parameters X_k and the Dirichlet distributions are conjugate to the multinomial model, using the update $X'_k \sim D(g_{k1} + J_{k1}, \dots, g_{kd} + J_{kd})$ for each row of X we have a Gibbs step (for details see [1]).

For the update of $\dot{Y}$ we can perform a Gibbs step drawing from the full conditional distribution $P(\dot{Y} | \dot{R}, \eta)$ using a standard forward-filtering-backward-sampling method (we refer to [2] for details).

For the update of Σ we generate a proposal Σ' by a joint normal random walk with step widths $\tau_{\xi_0}, \tau_{\underline{\sigma}}, \tau_\epsilon$, reflected at zero for $\underline{\sigma}$ and ϵ , i.e. $\xi'_0 = \xi_0 + \tau_{\xi_0} \psi_1$, $\underline{\sigma}' = |\underline{\sigma} + \tau_{\underline{\sigma}} \psi_2|$, $\epsilon' = |\epsilon + \tau_\epsilon \psi_3|$, where ψ_i are independent standard normal variates. This results in a Metropolis step with acceptance $\alpha_\Sigma = \min\{1, P(\dot{R} | M, \Sigma') P(\xi'_0) P(\underline{\sigma}') P(\epsilon') / (P(\dot{R} | M, \Sigma) P(\xi_0) P(\underline{\sigma}) P(\epsilon))\}$.

4.4 Numerical Results

We consider 50 samples of simulated prices ($d = 2$, $N = 2500$, $\Delta t = 1/250$, 600 000 iterations, 100 000 burn-in, $\lambda = 10$ fixed). In Table 1 the first row gives the exact parameters, and the second, third and fourth the mean, standard deviation, and root mean square error of the mean estimates. To avoid a bias the parameters for the prior distributions are chosen such that the mean values are the true ones; however, the variances (b : 36, $\underline{\sigma}$: 0.0225, ϵ : 400, ξ_0 : 0.2, X_{kl} : 0.007, $k \neq l$, where for X this corresponds to adding 50 observations from the true transition matrix to 2500 observations from the data given) are large, so there is little information a priori. While X is estimated, we also give the values for the corresponding Q . For simulated prices (with volatilities between 0.2 and roughly 0.3) going well with real financial data the results are quite satisfying. Note that the underestimation of Q is partly caused by the time discretization.

For an application to market data we refer to [2], where for daily data from the Austrian Stock Exchange Index plausible results are reported.

Table 1. Parameters and results for simulated data

	b	Q		X		$\underline{\sigma}$	ϵ	ξ_0
true	2.00 −2.00	−80.0 80.0	80.0 −80.0	0.76 0.24 0.24 0.76		0.200	30.0	0.000
mean	1.87 −1.89	−67.6 68.6	67.6 −68.6	0.79 0.21 0.21 0.79		0.200	25.3	0.002
std dev	0.24 0.26	13.7 12.6	13.7 12.6	0.03 0.03 0.03 0.03		0.006	5.8	0.039
RMSE	0.27 0.28	18.5 16.9	18.5 16.9	0.05 0.05 0.04 0.04		0.006	7.4	0.039

References

1. M. Hahn, S. Frühwirth-Schnatter, and J. Sass, *Markov chain Monte Carlo methods for parameter estimation in multidimensional continuous time Markov switching models*, preprint (2005).
2. M. Hahn, W. Putschögl, and J. Sass, *Portfolio optimization with non-constant volatility and partial information*, Brazilian Journal of Probability and Statistics (2006), to appear.
3. D. G. Hobson and L. C. G. Rogers, *Complete models with stochastic volatility*, Mathematical Finance **8** (1998), no. 1, 27–48.
4. E. Jacquier, N. G. Polson, and P. E. Rossi, *Bayesian analysis of stochastic volatility models*, Journal of Business & Economic Statistics **12** (1994), no. 4, 371–89.
5. A. Platania and L. C. G. Rogers, *Putting the Hobson-Rogers model to the test*, preprint (2005).

A Simulation Application for Predator-Prey Systems

Ulrike Leopold-Wildburger¹, Silja Meyer-Nieberg, Stefan Pickl², and
Jörg Schütze

¹ Karl Franzens Universität Graz, Austria
`ulrike.leopold@uni-graz.at`

² Universität der Bundeswehr München, Germany
`stefan.pickl@unibw.de`

Summary. We intend to optimize harvesting of two populations within a business cycle of an economy with appropriate means. The simulation model we have in mind concentrates on the problem of profit maximization within an interdependent system. Furthermore, we deal with the task of explaining phenomena of the typical behavior of the subjects while operating a complex system, e.g., a market, a company, ...

First studies of predator-prey systems give interesting insights into those subjects who behave optimally. In our model, maximal harvesting means letting the prey population grow first during the starting phase and taking constant quantities during the sustainability phase.

1 Introduction

The purpose of our model is to examine a dynamic system experimentally with the aim of optimizing harvesting within a sustainable environment. The complexity of the model (see also [6]) is based on the interdependencies of the two populations. Subjects are asked to deal with a simulation of increases and decreases in population numbers. The underlying Lotka-Volterra predator-prey experiment models an open, natural, discrete-state system. The main idea is based on [1].

2 The Model

The model underlying the experiment includes some fundamental assumptions listed here:

- isolated habitat,
- prey population grows logistically, predator population decreases exponentially,
- mutual influences are due to predator-prey relationship,

- prey population is diminished by predation,
- predation boosts predator population.

The mathematical background is given in the following way:
The basis for the formal model (1) is a system of two non-linear first-order differential equations:

$$\begin{aligned} dy/dt &= (-e + cx)y \\ dx/dt &= (a - by)x \end{aligned} \tag{1}$$

with $a, b, c, e > 0$. This system was first studied by [5] and [7]. The variables x and y are considered to be the populations of a species of prey and a species of predators, respectively. The coefficients, a , b , c , and e are positive constants.

Volterra considered the system as an attempt to explain the fluctuations in observed data on the catches of two species of fish in the Adriatic Sea. The populations of a species of small fish and a species of larger fish which feed on the small fish appear to fluctuate periodically, but out of phase.

After the large fish (the predators) kill off too many small fish (the prey) the large fish themselves begin to die out due to food shortage. This in turn gives the small fish population a chance to recover. But then the large fish population begins to boom again because of the abundance of food as shown by Figure 1. Our aim is to lead the system into an orbid where in harvesting of both populations can be maximized. Harvesting is done at the beginning of each period; the natural development determines the growth of both populations.

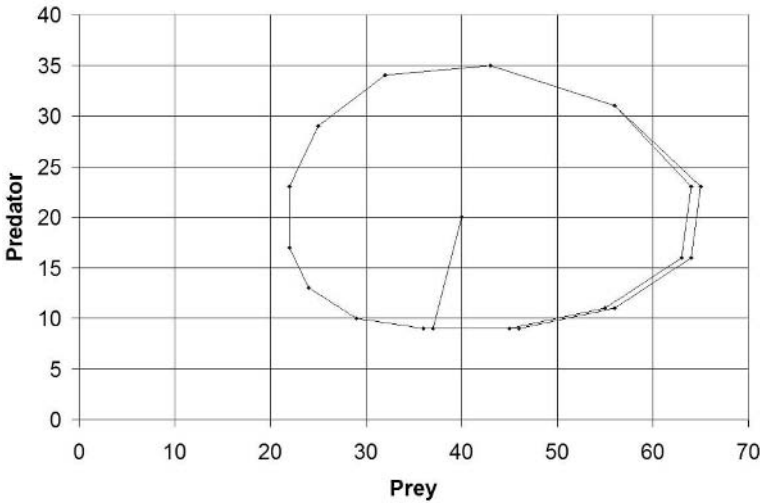


Fig. 1. Cycles of predators and prey

Although our experiment indirectly deals with an optimization (Goodwin's theory of the trade cycle, see [3]), we attack the problem in the field of biology.

Nevertheless, we intend to generalize the obtained results to global sustainable economic systems. Using different instruments, a biotope consisting of two populations represented by (1) should be brought into an optimal level in which harvesting can be optimized. A graphical representation of the system (1) can be given by the following diagram of Figure 2. The values of the constants are the following: $a = .6$, $b = .002$, $c = .001$, and $e = .4$.

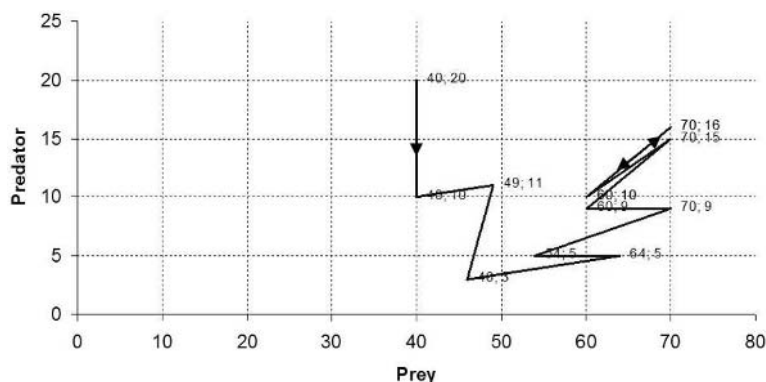


Fig. 2. Optimal strategy for the development of the species

Of course, some simplifications and idealizations make the system easier to handle:

- deterministic modeling as a closed system,
- discrete-time computation of a continuous-time system,
- continuous-state model of a discrete-state system.

Our methods for analysis are numerical integration methods and Lyapunov's indirect method to determine the stability of equilibria. The implementation is an application with elements as follows:

- predator-prey system with limited capacity,
- numerical integration using the Euler method and the fourth order Runge-Kutta method,
- graphical user interface,
- simulation component.

3 The Experimental Design

The experiment is established on a JAVA-based computer program SIMLVS by [2] which can be downloaded from the webpages of the authors.

The starting point of the experiment is a stable constellation with 20 predators and 40 prey. The challenge for the subjects is to maximize the harvest Q and the total sales S_{total} of the harvest for 42 periods. The prices P_{prey} and P_{pred} for the prey and for the predators are fixed. This can be written in the following way by Equation (2):

$$S_{\text{total}} = Q_{\text{pred}} \times P_{\text{pred}} + Q_{\text{prey}} \times P_{\text{prey}} \\ + \text{finalstock}_{\text{pred}} \times P_{\text{pred}} + \text{finalstock}_{\text{prey}} \times P_{\text{prey}} \quad (2)$$

Obviously, the total sales will be dependent on the prices, but also the optimality of the strategies will rely on the prices.

With 20 predators and 40 prey the populations will stay constant. In cases of more (less) than 40 prey the predators will increase (decrease). In cases of more (less) than 20 predators the prey will increase (decrease).

These are natural ceilings: The prey are limited to 70 due to the availability of food and, furthermore, the fishing quantities are limited to 10 per period and per population. At least 2 predators and 2 prey have to remain. This makes sense in order to ensure the system survives.

The subjects are asked to deal with the situation more than once; our suggestion is to play the game 10 times. The experiment was performed with 54 subjects. Each of them played the game 10 times. The subjects received financial rewards dependent on their performance. The price for harvesting a predator was fixed with 5 units each, the price for harvesting a prey was fixed with 3 units each.

4 Results

Obviously, the first goal of the subjects we are interested in is to bring the system into a maximal high level. These periods ($t(1) - t(5)$) can be called the starting phase.

Next, the level should be kept constant under the condition of harvesting as many predators and prey as possible. These periods from $t(6)$ onwards can be called sustainability phase.

Result 1 - Starting Phase

During the starting phase it is optimal to deal mainly with the predators with the aim of letting the prey grow. We classify the subjects as investors depending on how much they invest in prey. We group them into 3 classes: real investors, medium investors, and non-investors.

Our hypothesis is the following:

- H_0 : Real Investors achieve a far larger harvest and far more of the optimal total sales S_{total} than non-investors.

The variance analysis in Table 1 shows the significance of the results of the first experiments run with 54 subjects. Real investors reach 95% of the optimal total sales whereas non-investors reach only about 82%.

Table 1. Variance analysis for the investor behavior

Best Game – Investment Behavior	Sum of Squares	df	Mean Squares	F	Significance
Between Groups	0.132	2	0.066	8.119	0.001
Within Groups	0.423	52	0.008		
Total	0.555	54			

Result 2 - Sustainability Phase

During the sustainability phase it is optimal to deal mainly with constant harvesting for both populations. We classify the subjects as harvesters depending on how regularly they harvest prey and predator. We group them into 3 classes: Constant harvesters, medium harvesters, and jumping harvesters. Our hypothesis is the following:

H_0 : Constant harvesters achieve a far larger harvest and far more of the optimal total sales S_{total} than jumpers.

The variance analysis in Table 2 shows the significance of the results of the first experiments run with 54 subjects. Constant harvesters reach 97% of the optimal total sales whereas jumpers reach only about 85% during the sustainability phase.

Table 2. Variance analysis for the constant harvesting during the sustainability phase

Best Game – Investment Behavior	Sum of Squares	df	Mean Squares	F	Significance
Between Groups	0.134	2	0.067	8.306	0.001
Within Groups	0.421	52	0.008		
Total	0.555	54			

Finally, we are able to summarize that sustainable behavior optimizes the harvest and as a consequence also the total sales.

5 Conclusion

The economic analysis of investments with the aid of time-discrete dynamical systems and experimental methods leads to new insights in the area of resource-management. Many models in the area of environmental management concentrate either on mathematical or on political aspects. There is a lack of dynamical models,

which consider both and especially combine them with experimental issues. The software SIMLVS offers that possibility. It can be used to compare and classify optimal sustainable behavior. Furthermore, the results may be used to design future emission trading markets. The simulated examinations have a great impact on the design on those markets.

6 Acknowledgements

The authors thank Otwin Becker for the fruitful discussions and constructive suggestions. It was his idea to design such an experiment. Jürgen Fritsch and Stefan Siegelmann were very engaged in the development process of the user-friendly simulation tool SIMLVS.

References

1. Becker, O. and Leopold-Wildburger, U. (1995) Some new Lotka-Volterra Experiments. In: Kleinschmidt, J. et al. (eds) Operation Research Proceedings 1995, Springer Verlag, Berlin, p. 482-486.
2. Fritsch, J. and Siegelmann, S. (2006) Modellierung und Optimierung einer interaktiven Ressourcenplanung, Studienarbeit an der Universität der Bundeswehr, München.
3. Goodwin, R.M. (1982) Essays in economic dynamics, The Macmillan Press Ltd London, Basinstoke.
4. Krabs, W. and Pickl, S. (2003) Analysis, Controllability and Optimization of Time-Discrete Systems and Dynamical Games. Lecture Notes in Economics and Mathematical Systems, Springer Verlag, Heidelberg – New York.
5. Lotka, A.J. (1925) Elements of Physical Biology, New York.
6. Reichert, U. and Dörner, D. (1988) Heurismen beim Umgang mit einem einfachen dynamischen System. In: Sprache und Kognition, 7.Jg, Bd. 1, p. 12-24.
7. Volterra, V. (1927) Variations and Fluctuations in the Numbers of Coexisting Animal Species. In: Scudo, F. et al. (eds) 1978 Lecture Notes in Biomathematics: The Golden Age of Theoretical Ecology, p. 1923-1940.

Robustness of Econometric Variable Selection Methods

Bernd Brandl

University of Vienna, Austria
bernd.brandl@univie.ac.at

1 Introduction

Variable selection in cross-country growth regression models is currently a major open research topic and has inspired theoretical and empirical literature, see [6]. There are two categories of research problems that are intimately connected. The first problem is model uncertainty and the second is data heterogeneity. Recent literature aims to overcome the first problem by applying Bayesian Model Averaging (BMA) approaches in finding important, robust and significant variables to explain economic growth. While BMA offers an appealing approach to handle model uncertainty very little research has been undertaken to consider the problem of data heterogeneity. In this paper we analyze the issue of data heterogeneity on the basis of the exclusion of countries, i.e. we will take a closer look at the robustness of approaches when countries are eliminated from the data set. We will show that results of BMA are very sensitive to small variations in data. As an alternative to BMA in the cross-country growth regression debate we suggest the use of “classical” Bayesian Model Selection (BMS). We will argue that there is much in favor of BMS and will show that BMS is less sensitive in the identification of important, robust and significant variables when small variations in data are made. Our empirical results are undertaken on the most frequently used data set in the cross-country growth debate provided by [4].

2 The Cross-Country Growth Regression Debate

The problem encountered in the literature on the empirical validity of determinants of cross-country economic growth is that the number of explanatory variables is large compared to the number of observations. In traditional literature cross-country regressions of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon \quad (1)$$

were reported. Where y represents the vector of economic growth across countries, α is a constant, and $x_1, \dots, x_n$ are vectors of explanatory variables. In each paper

the set of explanatory variables differs according to different theoretical and technical considerations. For more details see also [4] and [2]. The aim of empirical research is to determine the importance of variables and specifications. In the context of economic growth, a regression with all variables that are argued to be important is simply not feasible for reasons of sample size and collinearity. The infeasibility inspired the development of data-driven approaches, see [1] and [3]. In recent literature the Bayesian Averaging of Classical Estimates (BACE) approach suggested by [4] appears to be most promising. Independent of using BMA or traditional regressions all literature makes the implicit assumption that the parameters are constant across countries, as stressed for example by [6]. This assumption is too strong as there are many examples where coefficients can be argued to vary extremely over countries. In the following we will observe the problem of parameter (or coefficient of variables) heterogeneity by a diagnostic analysis where we are excluding from the sample one country at a time. We will make a comparison of the robustness in the identification of important variables by applying the BACE approach and the BMS approach by optimizing the well-known Bayesian Information Criteria (BIC), proposed by [5].

3 Bayesian Model Averaging Versus Bayesian Model Selection

In the cross-country growth discussion the problem of selection and averaging arises when the relationship between y and a subset of x has to be modeled, but there is uncertainty about which subset to use. Model selection refers to using the data to select one model from the list of candidate models $M_1, \dots, M_K$, while model averaging refers to the process of estimating some quantity under each model M_j and then averaging the estimates according to how likely each model is. This means that model M_j is used together with the data to predict $\hat{y}_j$. This implies an overall prediction $\sum_{j=1}^k w_j \hat{y}_j$ where a particular weight w_j is used to express the probability that the model M_j generated the data. The BMA way is to compute the posterior probability for each model $P(M_j | Data)$. In the case of model selection, a model that maximizes $P(M_j | Data)$ is chosen. As our interest is not only in models but specifically on the importance of variables we will apply the frequently applied BACE approach to compute posterior probabilities and apply a genetic algorithm to solve the maximization problem. In contrast to other (pure) BMA approaches which demand the specification of a prior distribution for all parameters, the BACE approach requires only specification of the expected model size $\bar{k}$. The simple idea behind the BACE approach is to combine averaging of estimates across models with ordinary least squares (OLS) estimation. In the BACE approach, the posterior inclusion probability (PIP) provides information about the relevance of a variable i and the sum of the posterior probabilities of all models containing variable i is calculated via

$$P(\beta_i \neq 0 | y) = \sum_{j=1}^{2^K} P(M_j | y). \quad (2)$$

Where $P(M_j)$ is the posterior model probability, i.e. the probability distribution of model M_j given the data y . It is calculated as the proportional likelihood function corrected for the degrees of freedom

$$P(M_j | y) = \frac{P(M_j) T^{-\frac{k}{2}} SSE_j^{-\frac{T}{2}}}{\sum_{i=1}^{s^K} P(M_i) T^{-\frac{k_i}{2}} SSE_i^{-\frac{T}{2}}} \quad (3)$$

K denotes the overall number of possible variables, T defines the sample size and SSE denotes the sum of squared errors. As is well-known, the BIC, which we will use for the BMS and thus as our goal function in the optimization, uses the same approximation to the odds ratio to justify the criterion. The BIC is given by

$$BIC(k) = -2 \cdot \ln(L_n(k)) / n + k \cdot \ln(n) / n \quad (4)$$

Where $L_n(k)$ denotes the maximum likelihood of a model with k parameters based on a sample of size of n . However, both approaches are derived using Bayesian arguments and both approaches (in our application) have many similarities so that a comparison of their robustness and therefore ability as methods for the identification of important variables can be easily undertaken. We will discuss some robustness results of applying the two approaches with respect to country heterogeneity (i.e. when we drop only one country out of the sample).

4 Empirical Results

As a benchmark for the variable robustness of both approaches we use the results of both approaches applied on the full sample (i.e. including all countries). The results of the approaches can be seen in Table 1.

We will focus our comparison (i) on which variables are identified by the two approaches and how many other variables are found to be important when one country is dropped and (ii) on the number of variables that are missing compared to the result of the full sample. This comparison demands a similar model size so that according to [4] we will assume a model size of 7 variables and will use only the 7 most important variables that are identified by the approaches. In Fig. 1 and Fig. 2 we report the heterogeneity in the identification of important variables when dropping one country.

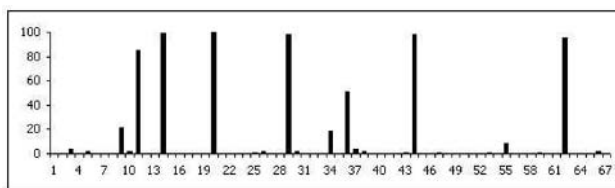


Fig. 1. Selection frequency of variables applying Bayesian Model Averaging

On the x-axis of Fig. 1 and Fig. 2 all variables used in the data set are listed. The ranking of variables is according to [4]. On the y-axis the percentage frequency of variables is given. When for a specific variable 100 percent is indicated, this variable is picked in every specification (when one country is dropped out of the sample). While we see for both approaches five variables that are always considered (and

Table 1. Optimal BIC results

Variable rank #	Variable	BACE 7	First best specification	Second best specification	Third best specification
-	Constant	0.0674*** (0.0139)	0.0722*** (0.0140)	0.0908*** (0.0183)	0.0670*** (0.0132)
3	Ethnolinguistic Fractionalization	-	-0.0103** (0.0046)	-	-
9	Fraction Confucian	-	-	-	0.0448*** (0.0163)
11	Population Density	8.51E-06***	8.87E-06***	9.78E-06***	6.28E-06***
14	Coastal in 1960s East Asian Dummy	(2.52E-06) 0.0197*** (0.0041)	(2.42E-06) 0.0202*** (0.0041)	(2.41E-06) 0.0186*** (0.0041)	(2.46E-06)
20	GDP in 1960 (log)	-0.0076*** (0.0020)	-0.0078*** (0.0019)	-0.0100*** (0.0023)	-0.0077*** (0.0019)
29	Investment Price	-8.12E-05*** (2.25E-05)	-8.81E-05*** (2.14E-05)	-8.09E-05*** (2.18E-05)	-9.12E-05*** (2.23E-05)
30	Latin American Dummy	-	-	-	-0.0151*** (0.0029)
36	Malaria Prevalence in 1960s	-0.0063 (0.0041)	-	-	-
44	Primary Schooling in 1960s	0.0285*** (0.0058)	0.0277*** (0.0057)	0.0299*** (0.0055)	0.0328*** (0.0058)
47	Political Rights	-	-	-0.0019** (0.0009)	-
55	African Dummy	-	-	-	-0.0182*** (0.0035)
62	Fraction of Tropical Area	-0.0142*** (0.0033)	-0.0152*** (0.0029)	-0.0167*** (0.0028)	-
Adjusted R-squared		0.7184	0.7271	0.7269	0.7161
Log likelihood		283.7604	285.1431	285.1018	283.3949
F-statistic		32.7075	34.1165	34.0738	32.3424
BIC		-6.0421	-6.0735	-6.0726	-6.0338
Frequency		-	67	1	7

Note: Numbers in parenthesis indicate standard errors. *** significant at 0.01, ** significant at 0.05. Column “Variable Rank” indicates the variable rank according to [4]. In column “BACE 7” the estimation results of using the 7 most important variables are shown. We restrict the number of variables to 7 as this is supposed to be a representative model size (for the justification see [4]). In columns “First best specification” to “Third best specification” the estimation results of the BIC optimization are shown. We report three versions as a stochastic algorithm is used to solve the optimization problem. In row “Frequency” we report the respective probabilities (in percent) of convergence to the specific solution. As can be seen convergence to the “First best specification” is most likely.

therefore are identified to be highly robust) we observe a higher heterogeneity for the BACE approach as several other variables (i.e. 18) are included occasionally. BMS is more stable in this respect as only 13 other variables are selected. We also look at the number of variables that are not included when one country is dropped

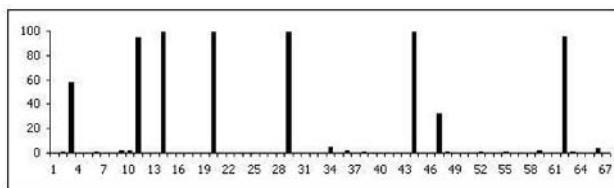


Fig. 2. Selection frequency of variables applying Bayesian Model Selection

out of the sample. The results for both approaches are given in Figure 3 and Figure 4.

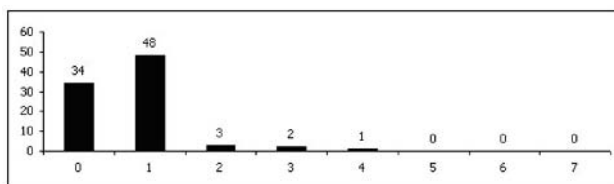


Fig. 3. Bayesian Model Averaging: Number of missing variables

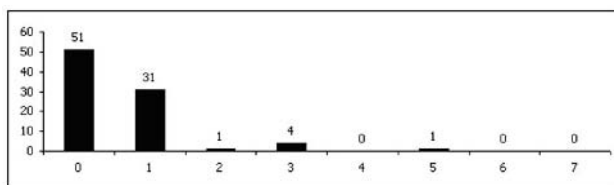


Fig. 4. Bayesian Model Selection: Number of missing variables

On the x-axis of the histograms in Fig. 3 and Fig. 4 the categories can be found which indicate how often variables are not included - compared to the 7 variables of the full sample. In the BACE approach for 34 specifications all variables are the same as in the benchmark model (this is indicated by a difference in the number of variables of 0). In 48 specifications only 1 variable differs from the benchmark. On average 0.727 variables are differing compared to the benchmark with a standard deviation of 0.739. The mode is 1, i.e. in 48 specifications 1 variable is different. The BMS approach results in 51 specifications in exactly the same specification. This is indicated by a difference in the number of variables of 0. In 31 specifications only 1 variable differs from the benchmark. We find no specification for which all the variables are different. On average 0.568 variables are differing compared to the benchmark with a standard deviation of 0.881. The mode deviation is 0 (i.e. in 51 specifications all variables are the same as in the benchmark specification). This, however, shows that BMS is more robust in the identification of variables than BMA.

5 Conclusions

Our empirical result showed that BMS is more robust in the identification of variables when one country is dropped out of the sample. As dropping one variable is a relatively small change in the data one can expect that nearly no changes in the importance of variables are found. It is surprising that BMS is more robust, especially when thinking about the fact that the results of BMS are based on the results of a stochastic optimization procedure. However, besides computational aspects of choosing between BMA and BMS there are other aspects in choosing a specific approach which we have not addressed. For example, BMA is a method for not discarding variables, as usually BMA does not lead to a parsimonious set of variables. BMS tends to exclude important variables. The higher robustness of BMS may be explained by the fact that BMA is not able to choose between variables that are collinear, and when there is a true data generating process then the results of BMA are biased as positive weights to wrong models are attached.

References

1. Hendry DF Krolzig HM (2005) The Properties of Automatic Gets Modeling. *Economic Journal* 115:32–61
2. Hoover KD, Perez SJ (2004) Truth and Robustness in Cross-Country Growth Regressions. *Oxford Bulletin of Economics and Statistics* 66:765–798
3. Phillips PCB (2005) Automated Discovery in Econometrics. *Econometric Theory* 21:3–20
4. Sala-i-Martin X, Doppelhofer G, Miller RI (2004) Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. *American Economic Review* 94:813–835
5. Schwarz G (1978) Estimating the Dimensions of a Model. *Annals of Statistics* 6:461–464
6. Temple J (2000) Growth Regressions and What the Textbooks Don't Tell You. *Bulletin of Economic Research* 52:181–205

Using Shadow Prices to Reveal Personal Preferences in a Two-Stage Assignment Problem

Anke Thede and Andreas Geyer-Schulz

LS Informationsdienste und elektronische Märkte, Universität Karlsruhe, Germany
thede@em.uni-karlsruhe.de
geyer-schulz@em.uni-karlsruhe.de

1 Introduction

Many events such as conferences or informational events at institutions consist of a number of single workshops among which participants can choose a limited number to visit. The workshop assignment problem is a linear optimization problem that finds a feasible assignment of participants to workshops maximizing the social welfare function expressed by preferences of participants for workshops. Several problem constraints reflect the structure and limitations of the event and define possible workshop combinations.

The optimization problem is formulated based on a priori preferences of participants over the workshops. Preferences constitute private information. For several reasons, preferences may be interrelated and can thus be influenced by the preferences of other participants and the optimal assignment resulting from them. The publication of the optimal assignment reveals additional information to the participants that may influence their initial preferences. As a result, the assignment is no longer optimal with respect to the modified preferences which has a negative effect on the satisfaction of the participants with the assigned workshops.

To overcome this problem, we introduce a second optimization stage. During the second stage, participants can use the additional information revealed by the publication of the first assignment to modify their preferences. Shadow prices are used to communicate the necessary information in an efficient way. Based on the new preference values and a slightly modified problem definition, possible changes in the current assignment are calculated. These changes are restricted in such a way that no participant may be assigned a combination that is worse compared to his currently assigned workshops. Simulation results show that the so-defined second stage can effectively be used to improve the assignment and that shadow prices are a good means for providing the necessary information to the participants.

2 The Workshop Assignment Problem

The workshop assignment problem is formulated as a binary, linear maximization problem. It allows for workshop capacity restrictions, time restrictions on workshops as well as participants and specification of workshop equivalence. Let $\mathbb{B} = \{0,1\}$. Let $T = \{1, \dots, \bar{t}\}$ denote the timeslots or indivisible time units of the event. Let $W = \{1, \dots, \bar{w}\}$ denote the set of workshops where each workshop takes place exactly once. Workshops taking place multiple times are modelled as several equivalent workshops by the equivalence relation $\sim$. Participants should attend at most one workshop of each equivalence class where $[w]_\sim$ is the equivalence class of workshop $w \in W$, and $W/\sim$ is the set of all such classes. The positive entries in the matrix $\mathbf{Q} \in \mathbb{B}^{\bar{w} \times \bar{w}}$ specify equivalent workshops. $P = \{1, \dots, \bar{p}\}$ is the set of the participants. The capacity limitations of the workshops are given by the vector $\mathbf{c} \in \mathbb{N}^{\bar{w}}$. Each participant visits a minimum number of workshops, given by the vector $\mathbf{m} \in \mathbb{N}^{\bar{p}}$. The matrix $\mathbf{F} \in \mathbb{B}^{\bar{w} \times \bar{t}}$ is an interval matrix where the 1-entries specify the timeslots occupied by each workshop. The presence of the participants in the timeslots is given by the matrix $\mathbf{B} \in \mathbb{B}^{\bar{p} \times \bar{t}}$. The participants' preferences for the workshops are given by the matrix $\mathbf{S} \in \mathbb{R}^{\bar{p} \times \bar{w}}$. Preferences are specified using a partial preference function. A default preference value is assigned to all workshops that are not explicitly given a preference value by a participant. Preference values smaller than the default preference mark workshops that the participant does not wish to attend whereas higher values stand for a positive preference. Preferences are chosen by the participants from a pool of available preference values. This pool may contain the same preference value more than once, workshops with the same preference assigned are considered to be equally desirable for the participant.

With $x_{pw} \in \mathbb{B}$ denoting the assignment of participant p to workshop w , the workshop assignment problem is formulated as follows.

$$\max \sum_{p \in P, w \in W} s_{pw} \cdot x_{pw} \quad (1)$$

subject to

$$\sum_{p \in P} x_{pw} \leq c_w \quad \forall w \in W \quad (2)$$

$$x_{pw} = 0 \quad \forall p \forall w (p \in P \wedge w \in W \wedge \exists t (t \in T \wedge f_{wt} = 1 \wedge b_{pt} = 0)) \quad (3)$$

$$\sum_{v \in [w]_\sim} x_{pv} \leq 1 \quad \forall p \forall w (p \in P \wedge w \in W \wedge |[w]_\sim| > 1) \quad (4)$$

$$\sum_{w \in W} x_{pw} \geq m_p \quad \forall p \in P \quad (5)$$

$$\sum_{w \in W: f_{wt}=1} x_{pw} \leq 1 \quad \forall p \in P, \forall t \in T \quad (6)$$

$$x_{pw} \in \{0,1\} \quad \forall p \in P, \forall w \in W \quad (7)$$

Condition (2) ensures that the capacity limitations are met, (3) prohibits to assign participants to workshops that take place while they are not present, (4) prevents the assignment of one participant to more than one workshop from an equivalence class, (5) makes sure that each participant visits his required minimum number of workshops and finally (6) ensures that no two workshops are assigned that overlap in time.

3 Preference Dependencies

Why and how do preferences of participants depend on each other? There are several possible causes for such dependencies.

As sociopsychological publications show, there are people that tend to compliant behavior concerning social interaction and others that are rather detached [5]. Compliant people seek the company and acceptance of others and are likely to follow the crowd. Detached people prefer to be independent and stay away from the crowd. These traits have been shown to influence social behavior and thus introduce an interdependence of preferences [3, 7, 1]. Compliant people prefer popular workshops that receive high preference values of many other participants whereas detached people tend to the opposite type of workshops.

Social groups also affect preferences. Experiences have shown that groups of friends often like to visit the same workshops together, in this case the preferences of the social group members are interrelated. At the same time it is possible that a group of people prefers to get a broad overview of the workshops offered and therefore its members try to visit different workshops.

Another influence on the preferences is given by the partial preference function in the situation that a participant is assigned to one or more workshops with the default preference assigned. In this case the participant might want to refine the stated preferences to distinguish between those workshops that were a priori left unclassified.

4 Adding a Second Stage

Multi-stage procedures are already known from auction design [2, 4, 6]. Multiple stages help to publish important information during the auctioning process that can lead to a more efficient outcome than the one-stage counterparts, especially when personal values or preferences are interrelated and influence each other. Contrary to the workshop assignment problem, though, auction data is not very complex and the mere publication of all data is feasible. In our case, the problem data is composed of the preferences of the participants as well as the mechanism that calculates the assignment. As this information is too large and too complicated to be processed by the participants information has to be reduced to a smaller quantity that still conveys the most important information needed to update the preferences accordingly. For compliant and detached participants, the information needed is basically the popularity of the workshops. For social groups the assignment of group members can easily be communicated within the group, again the popularity of workshops is important to establish which workshops are the most promising exchange partners. It will be rather difficult to exchange a current place with a place in a popular workshop whereas there is a higher chance of successfully changing into a workshop that is less popular, especially less popular than the workshop of which a place is freed.

The popularity of workshops is given by the scarcity of the workshop as a resource. The scarcity is defined by the demand expressed by the participants' preferences. The shadow price is a very good measure to reflect this property. It corresponds to the increase in preferences for assigned workshops that would result from adding one additional place to that workshop. It reflects the additional, unsatisfied demand for that workshop as well as all other resulting improvements that would be generated for other participants. Moreover, the shadow price is zero for workshops with free places or for which there is no excess demand. These are the workshops that are most promising for a change of places. As the shadow price is just a single numeric value it allows for an ordering of workshops and gives participants an easy overview over the most promising possibilities for an exchange of

preferences. In fact, as we have a partial preference function with a limited number of possible preferences each preference modification consists in an exchange of preferences such that the total number of assigned preferences cannot change.

In order to offer incentives for participating or at least accepting the adoption of a second optimization stage it is very important to ensure for each new assignment that the quality of the assignments does not deteriorate for any participant. That is, the solution space of the new assignment has to be restricted to the part that is pareto efficient with respect to the currently valid assignment. Only those solutions that guarantee the same level of assignment to each participant are considered to be valid solutions of the second stage. Although, this guarantee does not apply to those workshops that are assigned beyond the requested minimum m_p of assigned workshops for each participant $p \in P$, so only the highest m_p preferences are considered.

To formulate the optimization problem of the second stage, the following data structures and functions are needed. The function $S^{exch}(p, w) \in \mathbb{R}$ is equal to s_{pv} if the preferences of w and v are exchanged by participant p , or s_{pw} otherwise. A workshop has at most one exchange partner. $min\_s(p, \hat{s}) = \{w \in W : \max(s_{pw}, S^{exch}(p, w)) \geq \hat{s}\}$ is the set of all workshops w with an original or exchanged preference of p of at least $\hat{s}$. The function $visit\_top(p)$ results in a set of size m_p of those currently assigned workshops with the highest preferences. If a participant is assigned exactly m_p workshops then $visit\_top()$ just contains all assigned workshops. The function $cnt(p, w) = |\{v \in W : z_{pv} = 1, s_{pv} \geq s_{pw}\}|$ gives the number of currently assigned workshops with a preference of at least s_{pw} . With $\mathbf{Z} \in \mathbb{B}^{\mathcal{P} \times \mathcal{W}}$ denoting the currently valid assignment, the new optimization problem is given as follows:

$$\max \sum_{p \in P, w \in W} S^{exch}(p, w) \cdot x_{pw} \quad (8)$$

subject to constraints (2) to (7) and

$$\sum_{v \in min\_s(p, s_{pw})} x_{pv} \geq cnt(p, w) \quad \forall p \in P, \forall w \in visit\_top(p) \quad (9)$$

The given formulation of the pareto-efficiency constraint (9) results in constraints with all factors of the problem variables equal to one. As this is also the case for all other constraints the problem structure is kept quite simple. On the other hand it produces m_p lines (or less) for each participant p . Other formulations are possible, e.g. one that compares the sum of the preferences of the assigned workshops with the corresponding sum of the underlying, current assignment.

5 Results and Conclusion

To show that the so defined two-stage procedure actually helps to improve the quality of the assignment a simulation was conducted. Assignment problems were generated based on a range of variable values that were found to be most representative for practical problems, with up to 100 workshops, up to 16 time slots and up to 1000 participants. The resulting problems had up to 113,500 variables and up to 22,600 conditions. Two types of preference interrelations were modelled, the first of which are detached participants that prefer less popular workshops and the second of which are social groups of size two, i.e. pairs of friends that desire to visit the same

workshops. The simulation allowed for each participant or pair of friends to issue exchange requests at most once, this reflects the limited patience of participants and transaction costs involved with active participation in the second stage. After each new request to exchange preferences of workshops the second-stage optimization problem was generated based on all pending requests and solved. The new solution was analysed and compared with the currently valid assignment. For each participant with an active request, for whom a modification of the assignment was detected all exchange requests were deleted and the exchange of preferences was made effective. For participants with pending requests but without any change in assignments the exchange requests were kept for subsequent rounds.

The results for detached participants are shown in table 1. The values shown are averaged over the 100 problems generated. There were 102 detached participants (a ratio of 0.3 of all participants) with 140 assigned workshops. The second stage raised the number of workshops with zero shadow price (SP0, most preferred by detached participants) by 38.14 which corresponds to an increase in the ratio of SP0 workshops from 0.52 to 0.79 of all workshops of detached participants. The last four columns show the ratios of participants for which exchange requests were successfully generated and if and when these resulted in a change of workshop assignment. For 66% of the participants, an exchange request was generated and resulted in a modified assignment (51% immediately, 15% later), while for 13% the generated requests did not result in any assignment modification. For 21% of the participants a sensible exchange request could not be generated at all.

Table 1. Results for detached participants

#det- ached	#work- shops	#new ws with SP0	ratio ws SP0 before	ratio ws SP0 after	exch. immed.	exch. later	no exch.	no request
102	140	38.14	0.52	0.79	0.51	0.15	0.13	0.21

Table 2. Results for friends

#pairs	requests (all/left)	same same	same diff	diff same	diff diff	#diff ws (before/after)
50.85	14.04/1.59	43.01 84.58%	0.64 1.26%	5.64 11.09%	1.56 3.07%	7.42/2.22

The results for friends are given in table 2, again averaged over a total of 100 generated problems. 30% of the participants were randomly selected to have a friend, which resulted in an average of 50.85 friend pairs. 14.04 exchange requests were made by these friends of which 1.59 remained without any effect until the end of the simulation. Columns 3 to 6 show how many friends visit the same or at least one different workshop before and after the simulation. In fact, as friends assigned the same a priori preferences to workshops a high percentage of more than 85% already visit the same workshops in the initial assignment. A small percentage of these, namely 1.26% of all friends, end up after the simulation with a worse situation, as they get assigned to different workshops. This is possible because pareto-efficiency is not guaranteed for social groups. But, for 11% the second stage improves the

situation, friends get assigned to the same workshops. 3% of the friends start and end with at least one different workshop. Over all friends, the number of different workshops is decreased by the second stage from 7.42 to 2.22 workshops.

The results show that the second stage successfully helps to improve the assignment for different kinds of preference interrelations while avoiding to deteriorate the situation for those participants not concerned by the stage. The information conveyed by the shadow prices is sufficient for the participants to choose promising workshops for the exchange of preferences that result in the algorithm successfully assigning them to the desired workshops.

References

1. Manfred Amelang and Dieter Bartussek. *Differentielle Psychologie und Persönlichkeitsforschung*. Kohlhammer, 1997.
2. Steven J. Brams and Alan D. Taylor. *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press, Cambridge, 1996.
3. Joel B. Cohen. An interpersonal orientation to the study of consumer behavior. *Journal of Marketing Research*, 4(3):270–278, 1967.
4. Yuzo Fujishima, David McAdams, and Yoav Shoham. Speeding up ascending-bid auctions. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 554–563. Morgan Kaufmann Publishers Inc., 1999.
5. Karen Horney. *Our Inner Conflicts*. W.W. Norton & Company Ltd, 1945.
6. Motty Perry, Elmar Wolfstetter, and Shmuel Zamir. A sealed-bid auction that matches the english auction. *Games and Economic Behavior*, 33:265–273, 2000.
7. Morris Rosenberg. *Occupations and Values*. Glencoe, Ill. Free Press, 1957.

Part VIII

Energy and Environment

Scheduling of Electrical Household Appliances with Price Signals

Anke Eßer¹, Andreas Kamper³, Markus Franke³, Dominik Möst¹, and Otto Rentz¹

¹ Institute for Industrial Production, Universität Karlsruhe, Germany
`anke.esser@wiwi.uni-karlsruhe.de`

² Institute for Applied Informatics and Formal Description Methods, Universität Karlsruhe, Germany

³ Institute for Information Systems and Management, Universität Karlsruhe, Germany

1 Introduction

Due to the increasing competition in liberalized electricity markets, a successful customer retention as well as a cost efficient allocation of electric energy become more and more important. Therefore, new, innovative strategies are sought, which promise on the one hand a long-term customer retention and assure, on the other hand, a more cost-efficient provision of electric energy.

Against the background of the rising cost awareness of customers, caused by the latest increases in electricity prices, price signals could be considered as an adequate instrument for this purpose (cf. e.g. [1], [2]). In this context, price signals can be defined as time-dependent rates for electricity, which are adjusted by power companies depending on the current or expected load.

The objective of this paper is to present a model development to analyse the influence of price signals on power demand as well as the results of its application to a rural area.⁴ For this purpose, it was assumed that in a model area consisting of about 1000 households, hourly prices are transmitted to the customers by their power suppliers one day in advance. Based on these prices, the customers decide whether to use their appliances at the usual time, or to reschedule the usage.

2 Modelling Approach for the Optimization of Households' Power Demand

In this paper an adequate modelling approach to analyse the influence of price signals on households' power demand is presented. It is assumed that households

⁴ The paper presents inter alia an abstract of the project "Analysis of the effect of price signals of the load behaviour of selected consumers", which has been conducted on behalf of the EnBW AG (cf. [2]). Besides, the modelling approach is enhanced in the project SESAM, supported by the BMBF (cf. e.g. [3], [4]).

minimize the costs and therefore react on the price signals by rescheduling the use of their electric household appliances. To begin with, all big household appliances are considered, that is washing machines, dryers, dish washers, and cooking appliances. It is assumed that all appliances are switched on and off manually. Besides, an automatic load management system for refrigerators and freezers is assumed. In the following section the mathematical description of this problem will be given. After this, some of the most essential aspects of the modelling approach will be addressed, that is the modelling of the households' individual load and preference curves as well as the time-dependent electricity tariffs.

2.1 Mathematical Description of the Optimization Problem

When households decide about the scheduling of their household appliances, they make a trade-off between the savings they might realise in rescheduling their appliances and the loss in comfort. In the optimization problem this consideration is realised via a weighted sum of preference and price. Consequently the objective function of the optimization problem consists of two parts (for nomenclature please cf. Table 1):

$$\text{Min} \sum_{i \in A} \sum_{j=t_i}^{t_i + k_v(i) + k_s(t_i, i)} (p_{el}(j) \cdot (d_v(i) + d_s(t_i, i)) \cdot \alpha + r(i, t) \cdot (1 - \alpha)) \quad (1)$$

The first part is the cost related part, where the time-dependent electricity tariff $p_{el}(j)$ is multiplied by the power demand of the appliance. This power demand consists on the one hand of the regular demand of the appliance $d_v(i)$. Besides an additional demand $d_s(t_i, i)$ might occur, if the appliance is rescheduled. The second part of the objective function represents the preferences of the household, where $r(i, t)$ expresses the preference of the household to schedule an appliance i in t . The shape of $r(i, t)$ will be described in section 2.3. The weight of the price is α , the weight of the preference is $(1 - \alpha)$. All considered appliances are subject to the following constraints:

1. All household appliances have to be scheduled.

$$\forall i \in A \quad g_i \leq t_i \leq e_i \quad (2)$$

2. Temporal restrictions are considered.

$$\forall i \in A \quad e_i + k_v(i) + k_s(t_i, i) \leq t_i \quad (3)$$

3. Dependencies between individual appliances are considered. For an appliance $j \in A$ which has to be scheduled after an appliance $i \in A$ applies:

$$i \xrightarrow{\text{before}} j \quad \Rightarrow \quad t_i + k_v(i) + k_s(t_i, i) < t_j \quad (4)$$

Besides, for the automatic scheduling of refrigerators and freezers applies:

1. The output per timeslot is limited to the capacity of the appliance.

$$0 \leq b_{cold}(t) \leq y(t) \cdot b_{KG}^{max}(t) \quad (5)$$

2. The daily requirement in cold, expressed in electrical energy, which is needed to supply the cold, has to be covered.

$$\sum_t b_{KG}(t) = D_{cold} \quad (6)$$

3. The storable cold is limited.

$$c_{cold}(t) \leq c_{cold}^{max} \quad (7)$$

4. The cold stored in the cooling device equals the cold of the previous period plus the cooling output of this period minus the demand.

$$c_{cold}(t) = c_{cold}(t-1) - \frac{D_{cold}}{24} + b_{KG}(t) \quad (8)$$

In the original optimization problem storage heatings as well as combined heat and power plants are treated, too. But since the full description of this problem would exceed the volume of this paper, all parts related to the heat and power production are omitted here (see herefore [2]).

Table 1. Used variables, function, and constants

Symbol	Measuring Unit	Description
t_i	-	current scheduling of the appliance j
$y(t)$	kWh	binary unit on / off
$b_{KG}(t)$	kWh	output of cooling device in t
$c_{cold}(t)$	kWh	stored cold in t
$d_v(i)$	kWh	power demand of the appliance i
$d_s(t_i, i)$	kWh	additional power demand of the appliance i when scheduled in timeslot t_i
D_{cold}	kWh	Demand in cold
A	-	set of all appliances
b_{max}^{KG}	kW	maximum capacity of the cooling device
c_{cold}^{max}	kWh	maximum amount of cold stored
$k_v(i)$	h	runtime of appliance i
$p_{el}[t]$	Ct/kWh	price signal in t
$r(i, t)$	-	preference function
$k_s(t_i, i)$	h	additional runtime of the appliance i when scheduled in timeslot t_i
g_i	-	earliest scheduling time possible of the appliance i
e_i	-	latest scheduling time possible of the appliance i

The optimization problem is a mixed integer linear problem. As solver, a version of the simplex-algorithm (CPLEX from ILOG) is used. Except for the binary variables, all discontinuous or nonlinear parts of the objective function have been replaced by piecewise linear functions. The scheduling of all household appliances has a maximum complexity of $O(m^n)$, where m is the number of time slots and n is the number of appliances. However, this complexity is only achieved if all appliances

are dependent on each other. If no dependencies exist, the complexity is reduced to $O(nm)$. In our case, only few dependencies exist, e.g. between washing machines and dryers. Besides, the number of considered household appliances is quite limited. Thus, the problem can be solved in an acceptable time (10 min per 1000 households and day). In the following, the most important input data will be described in brief.

2.2 Modelling the Individual Load Curves

To obtain results as high-grade as possible, an accurate representation of the load curves is indispensable. Therefore, a bottom-up approach was developed to create individual, realistic load curves for each household. In a first step, the households are defined and grouped by size and physical presence of the residents. Then, those electric appliances which can be rescheduled easily are identified. In total 45 different big household appliances and 15 cooling devices are considered. The distribution of the household appliances among the households is realised using data concerning the stock of household appliances and the degree of equipment of German households. In the last step of the modelling approach, the individual load curves of the electric appliances are summed up and all loads not modelled individually are aggregated in a base load curve. This curve was compared with the general VDEW load profiles and a good consistency was achieved.

2.3 Determination of the Individual Preference Curves

When scheduling their consumer loads, households make a trade-off between their temporal preferences and the time dependent costs of the use of the appliances. This trade-off is expressed in the objective function via a linear weighting of price and preference. Therefore, individual preference curves were constructed out of the aggregated load curves of the individual appliances and the revealed preferences of the household. In doing so, it is assumed, that in general the whole period determined by the load curve is available for scheduling the appliance. Nevertheless, the individually preferred scheduling times t_0 as well as their neighbouring time slots should be preferred - the latter, since generally it is easier to delay the utilization of an appliance by one hour than by half a day. Thus, the preference curve is defined as follows:

$$r(i, t) = \begin{cases} 1, 1 \cdot r'_{max} & , \text{ if } t = t_0 \\ \max(0, 8 \cdot r'_{max}, r(i, t_0)) & , \text{ if } t \in (t_0 - 1, t_0 + 1) \\ r'(i, t) & , \text{ otherwise} \end{cases}$$

Thereby $r(i, t)$ represents the preference of the household to schedule the appliance i in t and $r'(i, t)$ the values of the original curve. r'_{max} is the maximum of the original curve. Thus t_0 and its direct neighbours obtain high loadings: t_0 the highest value and its neighbours 80% of the highest value. For further information please compare [2].

2.4 Time-Dependent Electricity Tariffs

In our analysis, several different price models have been developed and tested. Initially diverse time depending tariffs with hourly changing electricity prices, based

on wholesale electricity prices have been used. Later on, a step tariff was developed that fulfilled much better the requirements of relevant monetary savings and load smoothing, especially with respect to a real life application, where simplicity plays a major role.

3 Selected Results

Using the modelling approach presented above, the influence of price signals on the electricity demand of households was analysed. A special focus lay on the global load smoothing potential as well as on the individual cost saving potential for households.

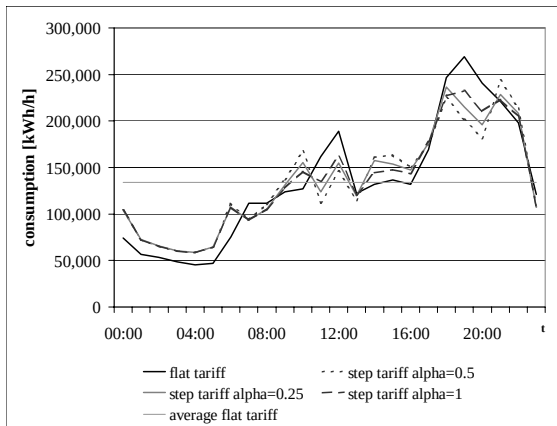


Fig. 1. Aggregated load curve of the model area

In Figure 1 the load smoothing for different values of the preference α is presented. A preference of 1 signifies that all big household appliances are scheduled according to the original preference of the household - only the cooling devices are scheduled on the basis of the price signal. It becomes apparent that in all optimized scenarios both the lunch peak and the evening peak are smoothened by more than 10%. Load rises mainly occur at night and in the afternoon. The former are caused by a rescheduling of the cooling devices, the latter by the big household appliances.

With an integration of a load management system for cooling devices, an average household in our model area realises savings of 15.03 Euros p.a., that is about 15,000 Euros per year for all households of the modell area. If a household begins to schedule a part of its big household appliances on the basis of price signals, further savings are possible, e.g. savings of 6.86 Euros per year with $\alpha=0.75$ and 13.12 Euros per year with $\alpha=0.5$.

4 Conclusion

Within this paper, a mixed-integer linear optimization approach to investigate if and to what extent price signals can be used to control power demand in an energy system is presented. Therefore it is assumed, that hourly prices are transmitted to the customers by their power suppliers one day in advance. Based on these prices, the customers decide whether to use their appliances at the usual time, or to reschedule the usage. They do so by weighting-up their temporal preferences and the time-dependent costs of the use of the appliance.

That way it can be shown that a high potential to shave off the lunch-time and evening peak exists in households and that price signals can be an adequate instrument to control the temporal distribution of power demand of households. In particular, to control cooling devices via price signals and load management systems seems remunerative. For the future, it is aspired to integrate the supply side in the modelling approach. In this context, especially the relevance of combined heat and power plants will be analysed more closely. Besides, based on the results of the modell, a field trial is planed to evaluate in reality the influence of price signals on households power demand.

References

1. Morovic T, Pilhar R, Möring-Hüser W (1998) Dynamische Stromtarife und Lastmanagement - Erfahrungen und Perspektiven. Technical report, Forschungsgesellschaft für umweltschonende Energieumwandlung und -nutzung mbH, Kiel
2. Rentz O, Eßer A, Möst D, Geyer-Schulz A, Franke M, Schmeck H, Kamper A (2006) Analyse der Auswirkungen von Strompreissignalen auf das Lastverhalten ausgewählter Verbraucher. Technical report, Universität Karlsruhe (TH)
3. Eßer A, Franke M, Kamper A (2006) Simulation und Einplanung dezentraler Energieerzeugungsanlagen. Technical report, Universität Karlsruhe (TH)
4. Eßer A, Franke M, Kamper A (2006) Modelle zur Steuerung, Planung und Optimierung in virtuellen Kraftwerken. Technical report, Universität Karlsruhe (TH)

Stochastic Optimization in Generation and Trading Planning

Thomas Hartmann, Boris Blaesig, Gerd Hinüber and
Hans-Jürgen Haubrich

Institute of Power Systems and Power Economics (IAEW), RWTH Aachen,
Germany

th@iaew.rwth-aachen.de

bb@iaew.rwth-aachen.de

gh@iaew.rwth-aachen.de

haubrich@iaew.rwth-aachen.de

1 Generation and Trading Planning

1.1 Motivation

In the process of the generation and trading planning, generation companies maximize the contribution margin, i.e. the difference of the revenues of energy trades and the costs for generating and purchasing electrical energy. The planning time horizon of this process covers a period of one month to one year. The complex planning task necessitates the application of computer-based tools [1]. Due to structural changes within the power industry, these tools have to be adapted continuously to the new conditions.

Nowadays these tools have to consider generation units, wholesale and reserve markets as well as fuel markets, limited transfer capacities and planning uncertainties.

1.2 Requirements for Planning Tools

Thermal power plants can be classified as single process units and combined-cycle gas turbines (CCGT), related to their different thermo-mechanical cycle. Each unit has to comply with technical restrictions like the minimum and maximum power output level, minimum up- and down-times as well as non-linear efficiency curves. Furthermore the accruing costs during start-up and normal operation have to be considered accordingly. Hydro power plants can be separated into storage and pump-storage as well as run-of-river power plants. Especially in mountainous regions hydro power plants are usually interconnected and characterized by interdependencies for the operation. The generation of electricity in hydro power plants is a non-linear function of the water flow and the hydraulic head.

Due to the deregulation of the electricity markets in Europe new power markets have been established. Electrical energy can be traded bilaterally (over-the-counter;

OTC) or at power exchanges. At the spot market electrical energy can be sold and purchased one day prior to delivery, whereas futures markets facilitate the trade of long-term contracts. The most liquid long-term products are base and peak which can be traded up to several years in advance. In addition, options on electrical energy provide opportunities of risk management in combination with futures.

Transmission system operators (TSO) have to procure reserve power in a tender to facilitate a secure network operation. For generation companies these newly established markets offer an alternative to wholesale markets. Hence, the decision has to be made whether to market the generation assets at the wholesale market or at markets for reserve.

Besides trading electrical energy, the trading of fuels, i.e. hard-coal and natural gas, gains in importance in the planning process of generation companies. Markets for hard-coal are already well established whereas the markets for natural gas emerge due to the recent deregulation of the gas sector. In analogy to the electricity markets spot and futures markets for gas emerge.

Driven by the Kyoto Protocol a reduction mechanism for greenhouse gas emissions has been established by means of a certificate trading approach. Fossil fuel fired power plants need to have sufficient allowances according to their emissions, which can be sold and purchased at the emission certificates market. Thus, the costs for the certificates have an influence on the specific fuel costs of a unit.

The interconnections between countries were originally not dimensioned for cross-border trading. The limited transfer capacities lead to restricted access to international markets. Hence, these limited capacities affect the generation and trading and have to be considered in the planning process.

The consideration of the components described above leads to a complex optimization problem. Because of the highly uncertain input data within the generation and trading planning, e.g. market prices and natural inflow of hydro power plants, stochastic optimization models and methods for risk management gain in importance.

The scenario analysis is an appropriate method to model these planning uncertainties in the generation and trading planning. The uncertainties are modeled by a scenario tree that starts with a deterministic root and then branches with increasing time to represent possible future developments [2].

In the following chapter an optimization model is presented that is designed as a planning toolbox in order to customize the model to a variety of present generation and trading planning tasks.

2 Optimization Model and Further Development

Considering the technical and economical properties of the system presented in section 1.2, the optimization problem that has to be solved is characterized by non-linearities, system and time spanning constraints as well as integer decisions. Therefore a multi-stage optimization model was developed to solve this problem with appropriate modeling accuracy. Fig. 1 gives an overview of the developed optimization toolbox.

In the optional first stage a linear approach determines the exchange schedule between different market areas with limited cross-border transfer capacities. Furthermore, this module can be used as a stand-alone tool to calculate the optimal

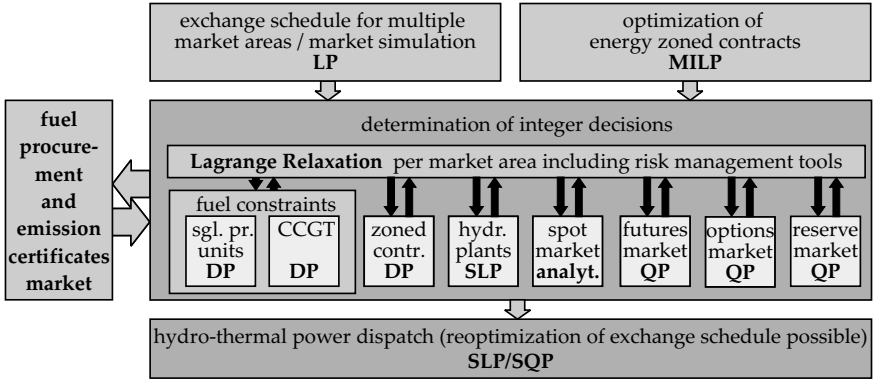


Fig. 1. Overview of the optimization toolbox

schedule for the power plants using linear programming (LP). Alternatively this stage optimizes zoned energy contracts and determines the optimal utilization of the contract by means of mixed-integer linear programming (MILP).

Using the exchange schedule or the optimal decisions for the zoned contracts respectively, the power plants are optimized against the different markets. Therefore the optimal generation schedule as well as the trading decisions are determined in a second stage. For modeling accuracy reasons, a decomposition approach is applied which divides the problem into subproblems on a system level. These subproblems are solved separately by means of the most appropriate algorithms.

Due to this iterative Lagrange Relaxation used in stage two, the system constraints, i.e. the energy and reserve balance, may be violated. Therefore a third stage performs a hydro-thermal power dispatch on the basis of the integer decisions determined in the second stage, optimizing the remaining continuous non-linear problem.

The objective function of the optimization problem is the maximization of the expected value of the contribution margin considering the willingness to take risks. Using weight factors the tool can trade off between the expected contribution margin and the Conditional Value at Risk as a risk measure [3].

The optimization tool is capable of optimizing a generation portfolio with several markets by means of appropriate algorithms. The toolbox can be customized depending on the required modeling accuracy as well as the admissible computation time of the tool to provide a high flexibility. In the following the presented components are explained in detail.

2.1 Linear Model and Exchange Schedule

The linear model determines the exchange schedule between market areas. In this stage, the generation portfolio and the relevant markets are considered, taking into account the limited transfer capacities and costs for cross-border trading. The cross-border exchange can either be modeled by consideration of NTC or PTDF alternatively, where the PTDF model represents a linearized power flow. The results are passed on to the second stage of the toolbox, but can alternatively be reoptimized

in the third stage. However, this module can be used as a stand-alone tool to solve large systems, e.g. for the purpose of market simulations, where a linear modeling accuracy is sufficient or a higher accuracy can not be handled within reasonable computation times.

2.2 Zoned Energy Contracts

In this stage the optimal utilization of zoned energy contracts, i.e. contracts in which the energy price depends on the quantity procured, are solved using mixed-integer linear programming. The contracts are optimized independently for each market area. The results of the optimal zones for each contract are assigned to all following stages.

2.3 Integer Decisions

The Lagrange Relaxation optimizes the generation portfolio against the markets using the exchange schedule or the optimal contract zones respectively determined in the previous stage. The overall problem for each market area is divided into its subproblems which are solved as independent problems using appropriate algorithms. In order to coordinate the system spanning constraints for electrical energy and reserve, Lagrange multipliers, that can be interpreted as price incentives for electrical energy as well as reserve power and energy, are passed on to the submodules. The results of the linear model in stage one can be used to initialize these price incentives.

The submodules optimize the specific problems of the system components and return the resulting generated energy as well as the reserve power and reserve energy separately for each reserve market back to the Lagrange coordinator. The price incentives are updated depending on the convergence of each system spanning constraint and the subproblems are reoptimized. Due to this iterative process, the system constraints converge until a sufficient level of accuracy or a maximum number of iterations is reached.

Considering the willingness to take risks in the objective function, a Lagrange multiplier interpreted as a risk avoidance incentive is defined and updated in every iteration of the Lagrange Relaxation to control the maximum allowed risk and the relationship between risk and contribution margin.

2.3.1 Generating Units

The thermal power plants can be classified as single process units and combined-cycle gas turbines (see section 1.2) and are optimized individually against the price and risk avoidance incentives. Due to the integer decisions of the thermal blocks, stochastic dynamic programming (DP) is used to determine the optimal schedule. Fuel constraints, e.g. take-or-pay contracts, can involve several thermal blocks and are therefore considered by an iterative loop including all relevant blocks. If the fuel constraints are violated, penalty costs will be introduced to meet the contract limits. As shown in fig. 1, different fuel markets and the emission certificates market can be considered [4]. The prices for these markets as well as the outage behavior of the power plants are modeled with stochastic data.

A hydraulic network model is developed for each interconnected group of hydro power plants, so that the interconnections between the reservoirs as well as the distribution of the limited available water in the planning horizon are considered. Since hydro power plants have negligible variable costs, the optimization model determines the optimal schedule considering the price incentives for electrical energy and reserve. Successive linear programming is used to incorporate the non-linearities in the model. The natural inflow to the reservoirs is modeled as a stochastic process.

2.3.2 Markets

Electrical energy is traded at different markets which are modeled as separate components. The hourly spot market can be solved analytically since the decisions for the hourly products can be determined independently. The high uncertainty of the price behavior necessitates a stochastic price model. The futures market, however, considers the price behavior for the time period in which the products can be traded. Due to the time spanning dependencies, a quadratic programming (QP) approach is selected to determine the optimal decisions for trading futures contracts. In analogy to the futures market, quadratic programming is also applied to the options market.

As shown in section 1.2, reserve markets offer new opportunities to generation companies and have to be considered in the model. Generation companies receive revenues for offering reserve power and for delivering the requested reserve energy. According to the regulations of the TSO, a degree of freedom is the distribution of reserve on available plants to provide the reserve capacity and to deliver the requested energy. The participation in the different markets with time spanning products are optimized by quadratic programming considering prices for reserve power and the expected requested energy. The delivery of reserve energy can be split up between the participating plants and is modeled by a stochastic time series for each market. The prices for reserve power as well as the requested energy are modeled as stochastic data.

In addition to the markets for electrical energy and reserve, it is possible to trade derivatives and insurances on correlated values, e.g. fuel or weather derivatives.

The emission certificates market allows generation companies to trade emission certificates for the thermal generation of electrical energy. Therefore, the costs for the emission certificates are modeled as additional variable costs for the thermal power plants depending on the fuel specific emissions.

2.4 Hydro-thermal Power Dispatch

Since it cannot be guaranteed that the system spanning constraints are fulfilled precisely in the iterative Lagrange Relaxation, the integer decisions derived in the second stage are used to set up a continuous problem which is solved in the hydro-thermal power dispatch. Because of the non-linear problem structure a successive linear (SLP) or successive quadratic programming (SQP) approach can be selected alternatively. Depending on the computation time and the selected modeling accuracy, this stage is solved separately for each market area or is formulated for all market areas. For the same reasons it is possible to optimize the whole stochastic problem in one step or separately for each time or scenario tree section.

The results of the hydro-thermal power dispatch provide the optimal generation and trading decisions as well as the distribution function of the contribution margin and time-integral energy quantities. With the application of the presented stochastic multi-stage optimization tool an appropriate model of the system components as well as the relevant uncertainties can be considered in the planning process.

References

1. Neus H., Schmöller H. K., Pribičević B., Flicke H.-P. (2002) Integrated Optimisation of Power Generation and Trading - Requirements and Practical Experience. In: Annual Report 2002 of the Institute of Power Systems and Power Economics of RWTH Aachen University, Aachener Beiträge zur Energieversorgung, Bd. 87, Klinkenberg, Aachen
2. Schmöller H. K., Hartmann Th., Kruck I., Haubrich H.-J. (2003) Modeling Power Price Uncertainty for Midterm Generation Planning, IEEE PowerTech Proceedings, Bologna
3. Blaesig B., Haubrich H.-J. (2005) Methods of Risk Management in the Generation and Trading Planning, IEEE PowerTech Proceedings, St. Petersburg
4. Neus H. (2003) Integrierte Planung von Brennstoffbeschaffung und Energieeinsatz zur Stromerzeugung. Aachener Beiträge zur Energieversorgung, Bd. 95, Klinkenberg, Aachen

Design of Electronic Waste Recycling System in China

Kejing Zhang, Daning Guo, Baoan Yang, and Fugen Song

Glorious Sun School of Business and Management, Donghua University, China

zhangkj@dhu.edu.cn

dnguo@dhu.edu.cn

bayang@dhu.edu.cn

fgsong@dhu.edu.cn

Summary. After years of rapid economic development, China is calling for development of a circular economy and resource recycling, to solve both the resource shortage problem and the environmental problems. Based on the analysis of the current status and the challenges of the electronic waste (e-waste) recycling in China, this paper seeks to design a financially feasible and environmentally safe e-waste recycling system. An integrated assessment model is proposed, in which the environmental, social and economic impact of recycling scenarios can be assessed simultaneously. The economic impact can be derived from an optimization model for the reverse logistics network.

1 Introduction

As a resource-poor country, China's rapid economic development has caused resource depletion and environmental problems. The conventional linear economic growth model, relying on large scale of resource consumption, has to be changed to resource-saving growth model, which focuses on resource recycling and recovery.

E-waste, comprising 8% of municipal waste, is one of the fast growing waste fractions [8]. The total volume of e-waste generated domestically in China is huge. More severe threat comes from the rapidly increasing amount of imported e-waste. More than 70 % of the e-waste of the United States goes to China [4].

E-waste refers to discarded appliances, such as TVs, PCs, air conditioners, washing machines, and refrigerators, as well as a variety of associated waste products, such as electrical wiring, printed wiring boards (PWBs), and batteries. E-waste contains five categories of materials: ferrous metals, non-ferrous metals, glass, plastics and others. Over 60 % of e-waste is composed of metals such as iron, copper, aluminium, gold and other metals.

The e-waste recycling is becoming a profitable business opportunity, in which valuable materials can be recovered and reused as a favorable resource for the economic development. 2.7% of the e-waste are pollutants including hazardous materials such as cadmium, mercury and lead. In the current e-waste recycling system, the

valuable components and materials of the appliances are extracted by manual disassembly and open incineration, while the remainder is dumped. Such unregulated and risky processing of e-waste has resulted in health problems and deterioration of air, water and land quality [7].

The objective of this paper is to design an e-waste recycling system based on EPR¹. An integrated evaluation model and an optimization model for the e-waste recycling network are proposed.

2 The Status and Problem of the Chinese E-Waste Recycling

2.1 Characteristics of China's E-Waste Recycling Practice

E-waste is considered as a resource and income generating. Many urban and rural poor make their living by recycling e-waste. E-waste recycling in the unregulated profit-driven informal sector has caused human and environmental risks. The fast increasing volume of e-waste, chiefly due to illegal import, has been beyond the existing processing capacity.

2.2 Challenges Facing China's E-Waste Recycling Industry

To solve the problems, one well-regulated and environmentally sound recycling system has to be established. Three major obstacles are:

1. How to get sufficient e-waste to recycle?
The formal recycling enterprises are competing for the e-waste stream with the extensive informal recyclers, which adapt well to the e-waste market. Buying e-wastes from the consumers means a financial burden on them.
2. How to finance the e-waste recycling system?
According to the WEEE ordinance, producers will be responsible for the take-back, recycling and disposal of e-waste. But the Chinese producers will be concerned about their competitiveness and profit.
3. Pressure from legal regulations
The products exported to EU have to meet the environmental requirements according to EU WEEE and RoHS Directives [4]. The Chinese regulations are far from elaborate and their enforcement is lax.

All these challenges are related to the informal recycling practice, which continues to prosper. The key issue is to incorporate it into the regulated system.

¹ EPR is defined as an environmental policy approach in which a producer's responsibility for a product is extended to the post-consumer stage of the product's life cycle, including its final disposal [5].

2.3 Joint Efforts to Establish a Sound E-Waste Recycling System

To establish a financially viable and environmentally sound e-waste recycling system, joint efforts should be taken from four levels: national level, provincial level, and enterprise level as well as community level.

On the national-level, strict legal regulation as guidance, financial support for related research and preferential policy for the fledging e-waste recycling sector are to be considered. Stricter and more demanding regulation is an essential factor that can make changes in the e-waste recycling industry [9].

At the provincial-level, local authority should issue legal regulation, especially formulate standards to encourage the development of e-waste recycling.

At the enterprise-level, the producers should take the financial and physical responsibility to take-back and recycle the e-waste and follow firm-based environmental standards. The recycling technology and know-how from industrialized countries can be transferred to help solve the Chinese e-waste recycling problems.

At the community-level, efforts should be put to raise the stakeholders' (collectors, recyclers and consumers) awareness of the importance to have the e-waste recycled in a manner that protects humans and environment.

3 Design of an E-Waste Recycling System in China

3.1 Legal Regulations Concerning E-Waste Recycling

To tackle the problem facing China's e-waste recycling, a series of regulations (see Table 1) have been issued. The WEEE ordinance, China's RoHS, and technical policy are applied to strengthen the earlier regulations such as "Notice on Strengthening the Environmental management of WEEE", and "Law on the Prevention of Environmental Pollution from Solid Waste" [7].

The regulations, if enforced strictly, will contribute to long-term social, environmental as well as financial sustainability. But meeting the requirements of the regulations means increased cost. The manufacturers are concerned about the effects of the regulations and the already small profit margin due to the severe competition.

3.2 EPR Policy in China

EPR policy is the crucial point of the e-waste regulations, and also a part of the effort of the government to set up a circular economy in the electronics industry [1]. From the experience of two successful EPR systems from Switzerland and Germany respectively, the following five elements of EPR system are to be specified in detail in designing China's EPR system: 1) Legal regulation; 2) Rate of return target; 3) System coverage; 4) Producer responsibility; 5) System financing. The EPR policy in China is explained in 3.3 with the help of two pilot programs.

Table 1. Three major e-waste related regulations in China

Legal Regulations	Purpose	Time of Issuance
Management measures for the prevention of pollution from electronic products (China's WEEE ordinance)	Restrictions on the use of hazardous substances; green design; provision of information on the components, hazardous substances, and recycling.	Issued on Feb 28, 2006 by MII (Ministry of Information), Effective from Mar 1, 2007;
Ordinance on the management of waste household electrical and electronic products recycling and disposal (China's RoHS)	Mandatory recycling of WEEE, based on extended producer responsibility; certification for 2nd hand appliances, and recycling enterprises.	Submitted for approval to the state council in early 2005; NDRC (National Development and Reform Commission)
Technical policy for WEEE pollution prevention	green product design; increase the reutilization rate and standard of e-waste recycling	Issued on Apr 6, 2004; SEPA (State Environmental Protection Administration)

3.3 Pilot Programs

As a further effort to determine the most suitable model for the Chinese e-waste management system, two national pilot programs have been initiated by NDRC [4, 8]. The city of Tsingtao and the province Zhejiang were selected to establish pilot e-waste recycling systems. In Table 2, the two pilot programs, representing two different models, are compared from eight aspects [7].

The current e-waste recycling system is subject to great uncertainty; many issues are under research; concrete strategies are to be developed and evaluated from economic, social, and environmental aspects.

4 An Integrated Assessment Model for the Recycling System

Based on the analysis of the recycling system and mathematical models² abroad, an integrated assessment model is developed, with environmental, social and economic impact of recycling practice under consideration [3]. Fig. 1 shows the framework of an e-waste recycling system and the methodologies used in the assessment model.

- The economic impact can be derived from an optimization model for the reverse logistics network.

² The objective of this optimization model for the e-waste recycling network is to minimize the total recycling cost (including fixed and transportation costs from collection sites to dismantling facility, further to recycling facility), processing cost at the facilities, and the fixed cost of operating the facilities [2, 6].

Table 2. Comparison of two pilot programs

	City of Tsingtao	Province of Zhejiang
Reason	Home to china's leading electronic appliance producers, such as Haier, Aucma and Haixin.	High market penetration and consumption rate of electronic products
Model	Producer-owned recycling	"Specialised" disposal plant
Responsibility	Brand specific; Individual producer responsibility;	Collective, regional system ZETC joint with DADI as PROs.
Coverage	Own products	TV, PC, air conditioners, refrigerators, washing machines and PWBs; Covering 11 cities;
Capacity	600,000 e-waste items per year;	800,000 items per year; First phase construction by end of 2006 with capacity of 200,000 items per year)
Investment	Total US\$ 10 million; US\$ 1.5 million by government.	Joint investment by ZETC and DADI; US\$ 12.5 million for facilities; US\$ 250 000 for establishment of R & D centre.
Recycling rate	Not given	Not specified
Legal regulation	National-level regulations	National-level regulations; Provincial-level regulations (provisional WEEE Management Measure, issued on Sep 14, 2005)

- Quantitative assessment of social impact and environmental impact can be done with the help of Process Analysis and MFA (Material Flow Analysis)
- The modeled strategies can be analyzed by MADM (Multi Attribute Decision Making) approaches with sensitivity analysis, which provides deep understanding of the problem. In the integrated assessment model, a group of criteria are chosen from life cycle perspective: cost, energy consumption, resource depletion, human toxicity, eco-toxicity etc.

5 Conclusions

China's rapid economic development has depleted natural resources and caused severe environmental problems. The recycling of e-waste can provide solutions to both problems and help build a resource-efficient and environmentally friendly society. Based on the analysis of the e-waste recycling industry, a financially viable and environmentally sound e-waste recycling system is proposed. For the evaluation of e-waste recycling strategies, an integrated assessment model is developed, in which the environmental, social and economic impact of recycling scenarios can be as-

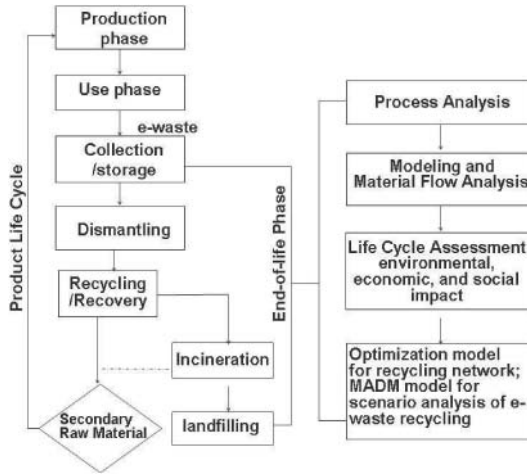


Fig. 1. The structure of e-waste recycling system and system assessment methods

sessed simultaneously. An optimization model for the recycling network provides data about the economic impact.

References

1. Feng Z (2004) Circular economy overview (in Chinese). People's publishing house, Beijing China
2. Fleischmann M (2001) Quantitative Models for Reverse Logistics. Lecture Notes in Economics and Mathematical Systems, vol. 501. Springer-Verlag, Berlin.
3. Geldermann J, Rentz O (2005) Multi-criteria Analysis for Technique Assessment. Journal of Industrial Ecology Vol.9, No.3: 127-142
4. Hicks C, Dietmar R, Eugster M (2005) The recycling and disposal of electrical and electronic waste in China-legislative and market responses. Environmental Impact Assessment Review 25: 459-471
5. OECD (2001) Extended producer responsibility: a guidance manual for governments. Paris
6. Spengler Th, Ploog M, Schroeter M (2003) Integrated planning of acquisition, disassembly and bulk recycling: a case study on electronic scrap recovery. OR Spectrum 25:413-442
7. Tao XZ, Zhong Y (2006) Analysis of the present Situation and Corresponding countermeasures of WEEE reverse logistics in China (in Chinese). Logistics SciTech 29(127): 48-50
8. Wei J, Li J, Wei H (2005) On reverse logistics model of WEEE recycle (in Chinese). Logistics Technology Vol.1: 28-30
9. Widmer R, Oswald-Krapf H, Sinha-Khetriwal D, Schnellmann M, Boeni H (2005) Global perspectives on e-waste. Environmental Impact Assessment Review 25: 436-458

A Coherent Spot/Forward Price Model with Regime-Switching

Lea Bloechlinger

Institute for Operations Research and Computational Finance, University of
St.Gallen, Switzerland
lea.bloechlinger@unisg.ch

1 Introduction

The challenge in modelling electricity prices is mainly caused by its non-storability. Spot prices are thus determined by the current demand/supply interaction, but hardly by expectations about the future. They show characteristics as mean-reversion, seasonal patterns, an immense volatility and spikes, which cannot be captured with standard stock market models. On contrary, there exists growing markets, where financial futures contracts are traded. These contracts are storable and show similar characteristics to other financial assets. In particular they feature a significant lower volatility than spot prices. Moreover, the volatility is decreasing in the time to maturity.

Hence, when modelling electricity prices one is actually confronted with two classes of prices, which show quite different characteristics. In the field of electricity price modelling there exist then also a number of models, which either focus on spot or on futures prices. However there are only few approaches, which try to capture the features of both prices in one model.

In this work we formulate and estimate a joint model, which describes spot and futures prices for electricity, taking the special characteristics of both prices into account.

2 Formulation of the Model

We consider an economy in which uncertainty is modelled by the filtered probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with $\mathcal{P}$ the real world probability measure. Events are revealed over time according to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t \in [0, T^*]}$, where T^* is a fixed time horizon, which limits the considered trading interval $[0, T^*]$. The market model contains electricity futures prices $F(t, T)$ of different maturities T as the prices of primary traded assets, where $0 < T < T^*$ and $t \leq T$. In addition it contains a riskless instrument. For simplicity of exposition, we assume that the risk-free interest rate is zero. The model also specifies the spot price process S_t , which is considered as a non-traded asset. We further assume that the market is free of arbitrage and that

there exists an equivalent probability measure $\mathcal{Q}$, under which futures prices are martingales.

The market security prices are assumed to be functions of k state variables $X_t = (X_t^1, \dots, X_t^k)'$. We consider two model specifications. The first model, called the basis model, is a two factor diffusion model, where X has the following dynamics under $\mathcal{P}$:

$$\begin{aligned} dX_t^1 &= (\kappa_1(X_t^2 - X_t^1))dt + \sigma_1 dW_t^1 \\ dX_t^2 &= \mu_2 dt + \sigma_2 dW_t^2. \end{aligned} \quad (1)$$

X^1 captures the short term movements of electricity prices. It is a mean-reverting process. The mean reversion level can be interpreted as the equilibrium price level, which is modeled as an arithmetic Brownian motion. The two Wiener processes W^1 and W^2 are independent. We assume further constant market prices of risk ϕ_1, ϕ_2 , such that the SDE's of the X^i under $\mathcal{Q}$ differ just by a constant reduction of the drift. The spot price S_t is composed of a seasonality factor $N(t) \equiv e^{n(t)}$ and the exponential of a weighted sum of the state variables:

$$S_t = N(t)e^{\gamma' X_t}. \quad (2)$$

In the basis model we choose $\gamma = (1, 0)'$. Assuming an arbitrage-free futures market, the futures price $F(t, T)$ must be a $\mathcal{Q}$ -martingale. Furthermore, the absence of arbitrage opportunity implies that it must converge towards the spot price at maturity. Thus the futures price is given as:

$$\begin{aligned} F(t, T) &= E_t^{\mathcal{Q}}[S_T] = N(T)E_t^{\mathcal{Q}}[e^{\gamma' X_T}] \\ &= N(T) \exp \left(\left(-\frac{\mu_2 - \phi_2}{\kappa_1} - \frac{\phi_1}{\kappa_1} \right) (1 - e^{-\kappa_1(T-t)}) + (\mu_2 - \phi_2)(T-t) \right. \\ &\quad \left. + \frac{\sigma_1^2 + \sigma_2^2}{4\kappa_1} (1 - e^{-2\kappa_1(T-t)}) + \frac{\sigma_2^2}{2}(T-t) - \frac{\sigma_2^2}{\kappa_1} (1 - e^{-\kappa_1(T-t)}) \right. \\ &\quad \left. + e^{-\kappa_1(T-t)} X_t^1 + (1 - e^{-\kappa_1(T-t)}) X_t^2 \right) \\ &= N(T) \exp(\alpha(t) + \beta(t) X_t). \end{aligned} \quad (3)$$

Applying Ito's-Lemma yields:

$$\frac{dF_t}{F} = e^{-\kappa_1(T-t)} \sigma_1 d\tilde{W}_t^1 + (1 - e^{-\kappa_1(T-t)}) \sigma_2 d\tilde{W}_t^2, \quad (4)$$

where $\tilde{W}^1$ and $\tilde{W}^2$ are Wiener processes under $\mathcal{Q}$. $\tilde{W}^1$ accounts for the fluctuations of futures prices with short maturities, whereas the second determines the long-term fluctuations. Thus there is no perfect correlation between futures prices of different maturities.

The second model we consider, the regime-switching model, is an extension of the first one. A third risk-factor X^3 is introduced. It is a pure jump process, which is modeled as a two state process with states J and K . The state variables X are then described by the following dynamics under $\mathcal{P}$:

$$\begin{aligned} dX_t^1 &= (\kappa_1(X_t^2 - X_t^1))dt + \sigma_1 dW_t^1 \\ dX_t^2 &= \mu_2 dt + \sigma_2 dW_t^2 \\ dX_t^3 &= 1_{X_t^3=0} d\nu_t^J - X_t^3 d\nu_t^K. \end{aligned} \quad (5)$$

$1_{X_t^3=0}$ denotes the indicator function. Thus, the process X^3 can jump into the spike state J only if it is in the normal state K , and it can jump back only, when it is in

the spike state. ν^J and ν^K are Poisson processes with intensities λ_J and λ_K . The jump size of ν^J , J , is normally distributed with mean μ_J and variance σ_J^2 . The jump size of ν^K is 1. The risk-neutral processes can be derived in the same way as in the basis model. For the third factor we assume that the risk premium is zero. In order to get tractable valuation formulas we further assume that W^1, W^2, ν^J, ν^K and J are mutually independent. The spike behavior is thus independent of the diffusive behavior, which means that the spot level has no influence on the spike occurrence.

The third factor can also be interpreted as two-state Markov regime-switching process. The according generator matrix, which describes the rate, at which the process jumps from state to state, is then given by

$$G = \begin{pmatrix} -\lambda_K & \lambda_K \\ \lambda_J & -\lambda_J \end{pmatrix}. \quad (6)$$

with K being the first and J being the second state. Building the matrix exponential $\exp(Gt)$ by diagonalizing G we get the transition matrix of the continuous Markov process:

$$P(\tau) = \begin{pmatrix} \frac{\lambda_N \exp(-(\lambda_N + \lambda_J)\tau) + \lambda_J}{\lambda_N + \lambda_J} & \frac{-\lambda_N \exp(-(\lambda_N + \lambda_J)\tau) + \lambda_N}{\lambda_N + \lambda_J} \\ \frac{-\lambda_J \exp(-(\lambda_N + \lambda_J)\tau) + \lambda_J}{\lambda_N + \lambda_J} & \frac{\lambda_J \exp(-(\lambda_N + \lambda_J)\tau) + \lambda_N}{\lambda_N + \lambda_J} \end{pmatrix}. \quad (7)$$

The spot price is now defined as in (2) with $\gamma = (1, 0, 1)'$ and the futures price is still given by the expected spot price at maturity under Q . Since we have assumed X^1 and X^3 to be independent we can write:

$$F(t, T) = \exp(n(T) + \alpha(t) + \beta(t)X_t^1)E_t^Q[\exp(X_T^3)], \quad (8)$$

where α and β have the same form as in (3). Taking into account the small data base for spike estimation, we make the simplifying assumption that the magnitude of the spikes is not random but constant of size μ_J . The state dependent expectations of $\exp(X_T^3)$ can then be written as:

$$\begin{aligned} E_t^Q[\exp(X_T^3)|X_t^3 = K] &= p_{KJ}(T-t)\exp(\mu_J) + p_{KK}(T-t) \\ E_t^Q[\exp(X_T^3)|X_t^3 = J] &= p_{JJ}(T-t)\exp(\mu_J) + p_{JK}(T-t). \end{aligned} \quad (9)$$

Following the presentations in [8], it can be shown that futures prices with relatively large maturities compared to the duration of a spike do not exhibit spikes. They can be given as a single expression:

$$F(t, T) = \exp(n(T) + \alpha(t) + \beta(t)X_t^1) (\pi_J \exp(\mu_J) + \pi_K) + O(e^{-\lambda_J \tau}), \quad (10)$$

where $\pi_N = \frac{\lambda_J}{\lambda_K + \lambda_J}$ and $\pi_J = \frac{\lambda_K}{\lambda_K + \lambda_J}$. The price given different states at time t differs then only by a term of order $O(e^{-\lambda_J \tau})$.

3 Model Estimation

To estimate the basis model described in (1) and (3), it is cast in *state space* form. The state space form consists of two parts. The *transition equation* describes the evolution of unobservable state variables $\xi \in R^k$ in discrete time:

$$\xi_t = c_t(\psi) + D_t(\psi)\xi_{t-1} + \eta_t(\psi). \quad (11)$$

ψ is the vector of the unknown parameters. The noise term η is assumed to be normally distributed and serially uncorrelated. The transition equation for the basis model is obtained solving the SDE's in (1). The second part, the *measurement equation*, relates the unobservable state variables to observable variables $y \in R^m$:

$$y_t = a_t(\psi) + B_t(\psi)\xi_t + \epsilon_t(\psi). \quad (12)$$

The noise term is again assumed to be normally distributed and serially uncorrelated. For the basis model the measurement equation can be derived taking the logarithm of (3). However, there occurs a problem. Futures prices with a delivery point T cannot be observed in the market. Only products with delivery periods of one month or more are traded. Hence, we define the future price with delivery period $[T_1, T_2]$ as $F(t, T_1, T_2) = \frac{1}{(T_2 - T_1)} \int_{T_1}^{T_2} F(t, s) ds$. To keep the linear relationship between state variables and the logarithm of futures prices, we substitute the geometric average for the arithmetic average.¹ Given the model in state space form Kalman filter [6] and maximum likelihood estimation method can be applied.²

For the estimation of the regime-switching model described in (5) and (10), we implement the Kim [9] filter. The Kim filter is an extension of the Kalman filter. It combines the Kalman filter with the Hamilton filter proposed for Markov-switching models³ and with a collapsing procedure. The Kim filter can be used for state space models, where the system matrices c_t , D_t in (11) and a_t , B_t in (12) depend on the regime R . The variable R_t , determining the current regime, is the outcome of a homogenous discrete-state Markov process with transition probability matrix P , $R^q \rightarrow R^q$. The regime-switching model can now be reformulated in such a way, that the third risk factor X_t^3 can be interpreted as the variable R_t , indicating in which regime the economy is. The transition equation remains then in both regimes the same as in the basis model. The only regime-dependent vector is $a_t(\psi)$, which must be chosen according to (9). The transition probability matrix P is given by (7).

For the empirical implementation we used spot and futures data quoted at the European Energy Exchange. The sample includes daily prices from January 2, 2003 to December 31, 2005. Only futures with base quality⁴ are included. The following maturities are considered: the three closest monthly products, the four closest quarterly and the two closest yearly products.

The main steps performing the estimation of the model parameters ψ are the following. First using a non-linear least-squares method we choose the initial guess for ψ by minimizing the differences between historical and model implied volatilities. Second, the following iterative procedure is started: for a specified value of ψ run the Kalman/Kim recursion over the whole time interval of the observation sample; after each complete run, evaluate the log-likelihood function and perform an optimization procedure⁵ in order to get new estimates for ψ ; as soon as the difference between

¹ Compare thereto [7], [1] or [2].

² For a technical description we refer to [4] or [5]. Applications in the field of energy modelling can for example be found in [10] or [11].

³ See [4] or [3].

⁴ Delivery from Monday to Sunday during all 24 hours.

⁵ We use a modified Downhill Simplex Optimization.

the value of the log-likelihood function of two consecutive iterations is smaller than the specified criterion, stop the iteration.

The estimation of the seasonality pattern $N(t)$ is done in a preceding step using regression models with temperature and calendar dummies as explaining variables. All risk premia are assumed to be zero.

4 Results

Table 1 lists the results of the parameter estimates.

Table 1. Estimated parameters of the basis and the regime-switching model.

Model Parameter	Description	Basis	RS
κ_1	short term mr-rate [y^{-1}]	19.89	3.34
σ_1	short term volatility [ann.]	3.7	0.58
μ_2	long term trend	0	0
σ_2	long term volatility [ann.]	0.13	0.13
$365\lambda_J$	avg. spike lifetime [d]	-	9.63
σ_2	avg. inter-spike time [d]	-	23.2
e^{μ_J}	spike multiplier	-	2.25
spot	std.dev. noise	0.1	0.21
month/1 month		0.10	0.21
month/2 months		0.14	0.10
month/3 months		0.14	0.10
quarter/1 quarter		0.09	0.06
quarter/2 quarters		0.09	0.09
quarter/3 quarters		0.09	0.09
quarter/4 quarters		0.09	0.09
year/1 year		0	0
year/2 years		0.04	0.03

The inclusion of the regime-switching has strong consequences on the parameters of the first risk-factors. This is caused by the fact that in the basis model the high spot price volatility as well as the much lower volatility of futures contracts with short time to delivery must be explained by the first risk factor. On the contrary, in the regime-switching model, the spot price volatility is also influenced by parameters of the third risk-factor, thus the volatility and as consequence also the mean reversion rate can be much lower. Figures 1 to 3 show some further results.

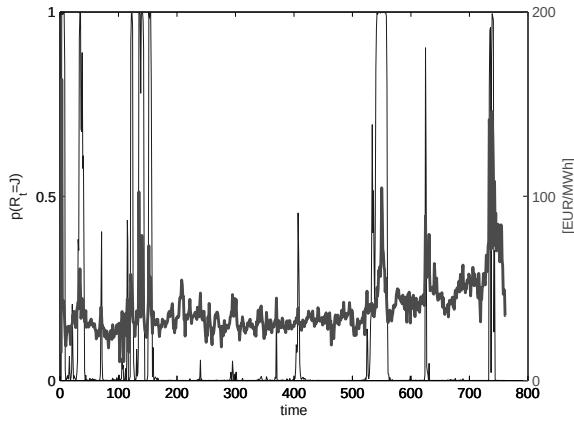


Fig. 1. Historical spot prices from 2003-2005 compared with the estimated probability of being in the spike regime. The estimation procedure identifies the different regimes well.

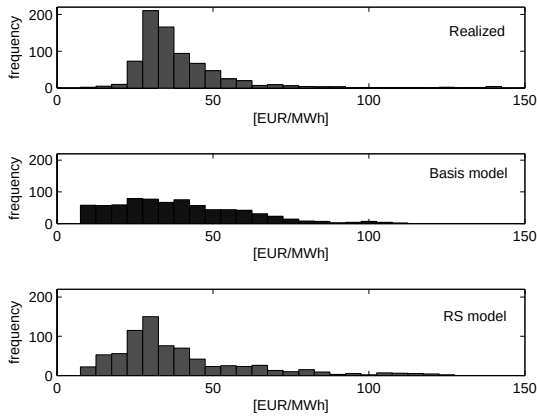


Fig. 2. Given the estimated parameters one spot price evolution is simulated for each model. The according histograms are then compared with the one of the historical spot prices. The regime-switching model provides a significant better fit.

5 Conclusion

We have presented two joint models for electricity spot and futures prices. The first model is a two-factor diffusion model. The second is enhanced by a Markov switching process. The models are estimated using filter technique and maximum likelihood method. The combination of the diffusion model with a Markov switching process improves the model fit significantly. In particular the distribution of the spot prices

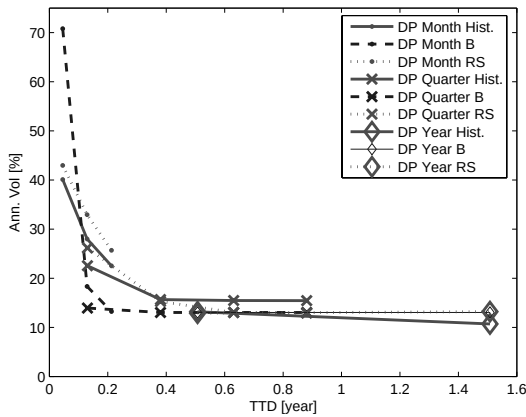


Fig. 3. Historical term structure of volatility compared with model implied term structures of volatility. The inclusion of the regime-switching improves the match clearly.

as well as the volatility term structure of the futures prices are better matched to the historical data.

References

1. P. Bjerksund, H. Rasmussen, and G. Stensland. *Valuation and risk management in the Nordic electricity market*. Working paper, Norwegian School of Economics and Business Administration, 2000.
2. M. Culot. *An integrated affine jump diffusion framework to manage power portfolios in a deregulated market*. PhD Thesis, Université de Louvain, 2003.
3. J. Hamilton. *Rational expectations econometric analysis of changes in regimes: An investigation of the term structure of interest rates*. *Journal of Economic Dynamics & Control*, 12:385–432, 1988.
4. J. Hamilton. *Time series analysis*. Princeton University Press, 1994.
5. A. Harvey. *Forecasting, structural time series models and the kalman filter*. Cambridge University Press, 1989.
6. R. Kalman. *A new approach to linear filtering and prediction problems*. *Journal of Basic Engineering*, 82:35–45, 1960.
7. A. Kemna and A. Vorst. *A pricing method for options based on average asset values*. *Journal of Banking and Finance*, 14:113–129, 1990.
8. V. Kholodnyi. *Modeling power forward prices for power with spikes: a non-markovian approach*. *Journal of Nonlinear Analysis*, 63:958–965, 2005.
9. C.-J. Kim. *Dynamic linear models with markov-switching*. *Journal of Econometrics*, 60:1–22, 1994.
10. M. Manoliu and S. Tompaidis. *Energy futures prices: Term structure models with kalman filter estimation*. *Appl. Mathematical Finance*, 9(1):21–43, 2002.

11. E. Schwartz and J. Smith. *Short-term variations and long-term dynamics in commodity prices*. Management Science, 46(7):893– 911, 2000.

Finance, Banking and Insurance

A Management Rule of Thumb in Property-Liability Insurance

Martin Eling, Thomas Parnitzke, and Hato Schmeiser

Institute of Insurance Economics, University of St. Gallen, Switzerland
martin.eling@unisg.ch

1 Introduction

Due to substantial changes in competition, capital market conditions, and supervisory frameworks, holistic analysis of an insurance company's assets and liabilities takes on special relevance. An important tool in this context is dynamic financial analysis (DFA). DFA is a systematic approach to financial modeling in which financial figures are projected under a variety of possible scenarios by showing how outcomes are affected by changing internal and/or external factors. The discussion in Europe about new risk-based capital standards (Solvency II project) and the development of International Financial Reporting Standards (IFRS), as well as expanding catastrophe claims, have made DFA an useful tool for cash flow projection and decision making, especially in the non-life and reinsurance businesses (for an overview, see [2]).

Some issues in implementing a DFA system have not been considered in the DFA literature yet. One of these is the integration of management strategies in DFA, which is the aim of this paper. Although addressing this issue has been recognized as necessary in the quest to improve DFA (see, e.g., [4], pp. 11–12; [2], p. 518), there is very little literature on the subject. [5] and [3] present theoretical discussions of the issue, but their aim is not to show practical implementations or to evaluate implications for DFA decision making.

The aim of this paper is to implement management strategies in DFA and study their effects on a property-liability insurer's risk and return position. We use performance measures that reflect both risk and return of these strategies in a multi-period context. Thus, the aim is to compare DFA with and without the implementation of specific management strategies so as to provide insight for an insurer's long-term planning process.

Our starting point will be a DFA framework that encompasses only the main elements of a property-liability insurance company (Sect. 2). Then, in Sect. 3, we develop typical management reactions to the company's financial situation. Sect. 4 contains a DFA simulation study to test the management strategies and examine their effects on risk and return. We conclude in Sect. 5.

In this paper, we present the findings for a basic model framework only. For an extended version of the model, and additional management strategies, the reader is

referred to [6], which also provides a detailed description of the applied performance measures as well as exhaustive numerical results and robustness tests.

2 Model Framework

With EC , we denote the equity capital of the insurance company and E stands for the company's earnings. For a time period $t \in T$, the following basic relation for development of the equity capital is obtained:

$$EC_t = EC_{t-1} + E_t. \quad (1)$$

The earnings E_t per period are comprised of the investment result I_t and the underwriting result U_t :

$$E_t = I_t + U_t. \quad (2)$$

On the asset side, high-risk and low-risk investments can be taken into account. High-risk investments typically consist of stocks or high-yield bonds; low-risk investments are usually government bonds or money market instruments. The portion of the high-risk investment in the time period t is given by α_{t-1} . The rate of return of the high-risk investment in t is denoted by r_{1t} and the return of the low-risk investment in t is given by r_{2t} . The rate of return of the company's investment portfolio in t , r_{pt} , is denoted by:

$$r_{pt} = \alpha_{t-1}r_{1t} + (1 - \alpha_{t-1})r_{2t}. \quad (3)$$

The company's investment result can be calculated by multiplying the portfolio return by the funds available for investment, i.e., the equity capital and the received premiums P_{t-1} less the acquisition expenses Ex_{t-1}^P :

$$I_t = r_{pt}(EC_{t-1} + P_{t-1} - Ex_{t-1}^P). \quad (4)$$

The other major portion of the insurance company's result is generated by the underwriting business. We denote β_{t-1} as the company's portion of the associated relevant market volume in t , assuming that $\beta = 1$ represents the entire underwriting market accessible to the insurance company. The volume of this underwriting market is denoted by MV . Hence, the total premium income can be obtained from:

$$P_{t-1} = \beta_{t-1}MV. \quad (5)$$

Claims are denoted by C ; expenses by Ex . Expenses consist of the acquisition costs Ex_{t-1}^P and claim settlement costs Ex_t^C . Acquisition expenses are calculated as a proportion of the premiums ($Ex_{t-1}^P = \gamma\beta_{t-1}MV$), while the claim settlement costs depend directly on the claims incurred ($Ex_t^C = \delta C_t$). Thus, we obtain the underwriting result with the relation:

$$U_t = P_{t-1} - C_t - Ex_{t-1}^P - Ex_t^C. \quad (6)$$

This model has two variables that management can change at the beginning of each period t —the proportion invested in the risky investment (α) and market share in the underwriting business (β).

3 Management Strategies

Management strategy arises out of a complex mixture of different business objectives (e.g., profit maximization, satisfaction of stakeholder demands, or maximization of the manager's own utility). Some strategies might require fast interventions in the event of a dangerous financial situation. Others will be less time-critical, e.g., growth targets or long-term profit maximization. Thus, management strategy can be various and multifaceted, depending on the actual situation of the enterprise and the aims of management.

A possible strategy is to reduce risk in a distressed situation so as to avoid insolvency. To define the solvency strategy, we set a level for equity capital such that when the capital falls below this level, a management reaction will be triggered. In the context of the European capital standards (Solvency I), the minimum capital required (MCR) can serve as this critical level. However, management will not wait to act until the equity capital falls under the MCR. Thus, we consider a safety loading of, e.g., 50% above this critical level. Following these considerations, the solvency strategy works as follows. For each point of time ($t = 1, \dots, T - 1$) we decrease α and β about 0.05 when equity capital is below the minimum capital required plus a safety loading of 50%.

An alternative is the growth strategy. This strategy combines the solvency strategy with a growth target for the underwriting business. Should the equity capital drop below the minimum capital required (MCR) plus the safety loading of 50%, the same rules apply as in the solvency strategy. If the equity capital is above the trigger, we assume a growth of 0.05 in β .

4 Numerical Example

We analyze a time period of $T = 5$ years. The parameters α and β can be changed in discrete steps of 0.05 within the range of 0 and 1 at the beginning of each year. The market volume (MV) of the underwriting market accessible to the insurance company is €200 million. Our model company has a portion of $\beta = 0.2$ in the market. The expenses, which are contingent on premiums written, are given by $Ex_{t-1}^P = 0.05\beta_{t-1}MV$.

Investment return is assumed to be normally distributed. The continuous rate of return has a mean of 5% (12%) and a standard deviation of 5% (20%) in the case of a low- (high-) risk investment. Hence, e.g., for the discrete risky rate of return, $r_{1t} = \exp(N(0.12, 0.20)) - 1$. The risk-free rate of return r_f is 3%. The claims C_t are modeled by using random numbers generated from a lognormal distribution with a mean of $0.85\beta_{t-1}MV$ and a standard deviation of $0.1\beta_{t-1}MV$. The expenses of claim settlement are determined by a 5% proportion of the random claim amount ($Ex_t^C = 0.05C_t$).

For the asset allocation we use data from the German regulatory authority (BaFin). German property-liability insurance companies typically invest about 40% of their funds in high-risk investments such as stocks and high-yield bonds, with the other 60% being invested in low-risk investments such as, e.g., government bonds or money market investments (see [1], Table 510). Thus we choose $\alpha = 0.4$ as the starting point for the asset allocation.

Calculation of the minimum capital required follows the Solvency I rules used in the European Union. The minimum capital thresholds based on premiums are 18% of the first €50 million and 16% above that amount. The margin based on claims, which is 26% on the first €35 million and 23% above that amount, is used if this sum exceeds the minimum equity capital requirements determined by the premium-based calculation. Following these rules we determine a minimum capital required of €8.84 million for $t = 0$ (i.e., maximum of $18\% \times €40$ million and $26\% \times €34$ million). To comply with the Solvency I rules, the insurance company is capitalized with €14 million in $t = 0$. This capitalization corresponds to an equity to premium ratio of 35%, which is a typical value for German property-liability insurance companies (see [1], Table 520).

The following simulation examples were calculated on the basis of a Latin-Hypercube simulation with 100,000 iterations (see [7]). The results are summarized in Table 1.

Table 1. Results

Measure		No strategy	Solvency strategy	Growth strategy
Return	$E(G)$	8.48	8.26	11.09
Risk	$\sigma(G)$	4.21	4.35	6.03
	RP	0.23%	0.11%	0.22%
Performance	P_1	2.01	1.90	1.84
	P_2	18.35	36.53	24.86

Notes:

$$\begin{aligned} E(G) &= \text{Expected gain per annum} = (E(EC_5) - EC_0)/5 \\ \sigma(G) &= \text{Standard deviation of gain per annum} = \sigma(EC_5)/5 \\ RP &= \text{Ruin probability} \\ P_1 &= (E(EC_5) - EC_0(1 + r_f)^5)/\sigma(EC_5) \\ P_2 &= (E(EC_5) - EC_0(1 + r_f)^5)/RP \end{aligned}$$

Looking at the simulation results in the case where no management strategy is applied, we find an expected gain ($E(G)$) of €8.48 million per year with a standard deviation($\sigma(G)$) of €4.21 million. The ruin probability (RP) is 0.23%, which is far below the ruin probability required by international regulatory authorities. For example, Switzerland requires a level of 0.40%. Thus our model company can be called safe.

The company becomes even safer if the solvency strategy is applied. While the return remains nearly unchanged (the expected gain decreases about 3% from €8.48 million to €8.26 million per year), we find much lower values for the ruin probability, which is now 0.11%. This is less than 48% of the value in the case where no management strategy is applied. Thus the solvency strategy is able to avoid most insolvencies without much effect on return. This also leads to much higher performance measures based on ruin probability (denoted as P_2 in Table 1) compared to the results in the “no strategy” case. The value of P_2 (36.53) is almost two times higher than that achieved with no strategy (18.35). Obviously, the solvency strategy

is capable of reducing downside risk effectively and thus provides valuable insolvency protection.

The growth strategy is more flexible than the solvency strategy. The expected gain per annum for this strategy amounts to €11.09 million, 31% above the €8.48 million obtained when no management strategy is applied. The ruin probability (0.22%) is slightly lower and thus P_2 is higher compared with the “no strategy” case. In comparison to the solvency strategy the growth strategy is suitable for those managers pursuing a higher return level and who are willing to accept a higher risk.

Interestingly, for both strategies risk is not reduced if both the positive and the negative deviations from the expected value are taken into account. For example the standard deviation of the solvency strategy is 3% higher than with no strategy (€4.35 million per year versus €4.21 million per year). This is because reducing the participation in insurance business and the fraction of risky investment changes the level of earnings among different time periods, which results in an increased standard deviation. Because of the higher standard deviation and the reduced return, the performance measure based on standard deviation (denoted as P_1 in Table 1) is slightly lower with the solvency strategy than in the “no strategy” case.

5 Conclusion

We implemented management strategies in DFA and studied their effects on a property-liability insurer's risk and return. We found that the solvency strategy, which reduces the volatility of investments and underwriting business in the event of a bad financial situation, is a reasonable strategy for managers who want to protect the company from becoming insolvent. Our numerical examples show that the ruin probability can be effectively delimited without affecting return too much. The growth strategy—combining the solvency strategy with a growth target in the insurance business—is an alternative for managers pursuing a higher return than offered by the solvency strategy and who are also willing to take higher risks.

References

1. BaFin (2005) Jahresbericht 2004, Statistik der Bundesanstalt für Finanzdienstleistungsaufsicht—Erstversicherungsunternehmen. Bundesanstalt für Finanzdienstleistungsaufsicht, www.bafin.de
2. Blum P, Dacorogna M (2004) DFA—Dynamic Financial Analysis. In: Teugels J, Sundt B (eds.) *Encyclopedia of Actuarial Science*. John Wiley & Sons, New York et al.
3. Brinkmann M, Gauß U, Heinke V (2005) Managementmodell. In: Lebensausschuss des DAV (ed.) *Stochastisches Unternehmensmodell für deutsche Lebensversicherungen*. Verlag Versicherungswirtschaft, Karlsruhe
4. D'Arcy SP, Gorvett RW, Herbers JA, Hettinger TE, Lehmann SG, Miller MJ (1997) Building a Public Access PC-Based DFA Model. *Casualty Actuarial Society Forum*, Summer(2)
5. Daykin CD, Pentikäinen T, Pesonen M (1994) *Practical Risk Theory for Actuaries*. Chapman & Hall, London

6. Eling M, Parnitzke T, Schmeiser H (2006) Management Strategies and Dynamic Financial Analysis. Working Paper, University of St. Gallen, St. Gallen
7. McKay M, Conover W, Beckman R (1979) A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* 21:239–245

Heuristic Optimization of Reinsurance Programs and Implications for Reinsurance Buyers

Andreas Mitschele^{1,2}, Ingo Oesterreicher¹, Frank Schlottmann², and Detlef Seese¹

¹ Institute AIFB, University of Karlsruhe, Germany

mitschele@aifb.uni-karlsruhe
deseese@aifb.uni-karlsruhe.de

IngoOest@gmx.de

² GILLARDON AG financial software, Bretten, Germany

frank.schlottmann@gillardon.de

1 Introduction

Reinsurance contracts represent a very important tool for insurance companies to manage their risk portfolio. In general, they are used if an insurer is not willing or not able to hold certain risk exposures or parts thereof on its own. There exist two main contract types to cede claims to a reinsurer, namely proportional and non-proportional ones. With the *quota share reinsurance*, a well-known variant of the former ones, a fixed percentage of the claim sizes is ceded to the reinsurance company. *Excess of loss* and *stop loss* are non-proportional types and the reinsurer is only liable to pay if certain losses are exceeded. In practice insurance companies usually place a number of different reinsurance contracts, a so-called *reinsurance program*.

While a substantial amount of research has been performed to optimize the terms and conditions of individual reinsurance contracts separately, the simultaneous optimization within a reinsurance program has hardly been analyzed to date and only few results exist so far³. Even though an integrated view of the reinsurance program is highly desirable, mathematical intractability, computational complexity and non-convexity of the objective functions have previously impeded the realization of this goal. In this contribution we therefore introduce a novel multi-objective approach to fulfil the task of optimizing reinsurance contracts simultaneously. Thus we minimize the expenses that come with contracting a number of different reinsurance protections and at the same time we minimize the retained risk after reinsurance.

³ Centeno [4] for instance focuses on combinations of quota share and excess of loss reinsurance, considering different cost calculation principles and risk measures. For a recent overview of current research see Verlaak/Beirlant [7] who also restrict their analysis to combinations of two reinsurance contracts.

2 Reinsurance Business and Optimization

In this section we briefly introduce some basic concepts from insurance mathematics⁴ and give an introduction to reinsurance contract optimization.

We consider three important types of reinsurance in this contribution. The simplest form is the *quota share* (QS), where a fixed relative amount $a \in [0; 1]$ of the claims and the premium income is ceded. In the other two types claims are cut at a positive line called priority. The *excess of loss* (XL) reinsurance cuts each individual claim at the priority R . The *stop loss* (SL) reinsurance cuts the sum of all claims in a year of business at the priority L . All contract types protect insurers from deviations in claim frequency and severity, however in complementary ways. Thus we contemplate combinations of two and three of these in our empirical investigation.

Reinsurance pricing is done via the expected value of the amount of claims ceded to the reinsurance company $\bar{S}$, i.e. $(1 + \lambda)E(\bar{S})$ ⁵. The risk of the retained net sum of claims $\underline{S}$ is assessed using three different measures, namely the variance, the common risk measure *Value at Risk* (VaR_α) and the more advanced *Conditional Value at Risk* (CVaR_α)⁶. VaR_α is denoted by the α -quantile of the distribution of $\underline{S}$ while CVaR_α represents the expectation of the conditional distribution below the α -quantile.

For reinsurance optimization we use a modified *Mean-Variance-Criterion* similar to the one used in [7]. As objective functions we choose the minimization of the cost of $\bar{S}$ and the minimization of the risk of $\underline{S}$ ⁷. Employing a heuristic approach has several advantages compared to the standard approach using the Lagrange method of multipliers. We do not need to transform to a single objective function. Moreover, our model incorporates discrete step sizes which is more realistic than a continuous framework. Centeno [4] proves that even in simple problem instances the feasible search space is not convex in general. Verlaak/Beirlant [7] state that the computation of optimal reinsurance contracts for a combination of XL and SL is an open problem⁸.

3 MOEA Approach

For the first time we treat the problem like a true multi-objective optimization problem and apply two state-of-the-art *multi-objective evolutionary algorithms* (MOEAs)⁹ to obtain risk/return efficient reinsurance contract allocations.

Hereby we pursue two main goals: Firstly, we intend to solve reinsurance optimization problems for cases which still pose open problems to researchers. Secondly, our aim is to find a converged and diverse but dense front of Pareto optimal solutions in a single run to enable the decision maker to choose the desired relationship

⁴ For a comprehensive introduction we refer to [1].

⁵ The factor λ describes the risk loading and depends on the reinsurance type.

⁶ Also known as *Expected Shortfall* (ES).

⁷ The distribution of the claims is based on real world data.

⁸ This problem of computational complexity is caused by the need to compute the distribution of the sums $\underline{S}$ and $\bar{S}$ numerically for each combination of R and L .

⁹ We compare NSGA-II [2] and ε -MOEA [3]. Details of chosen parameters cf. [6].

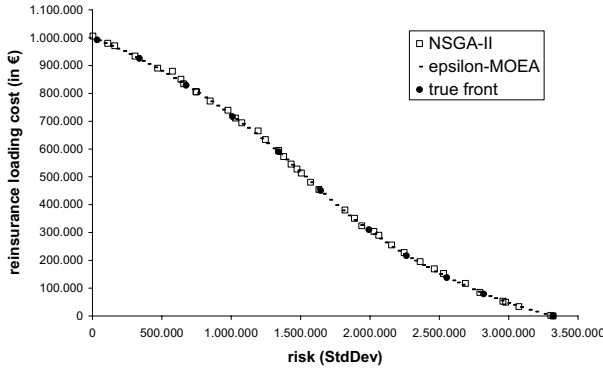


Fig. 1. Optimal fronts of the combination of QS and XL (Problem 1)

between risk and expense afterwards. A high density of the front is then used to deduce information about the structure of the reinsurance variables in optimal contract compositions.

4 Results and Implications for Reinsurance Buyers

Before analyzing more difficult issues we test the applicability of MOEAs for this problem class by considering a combination of quota share and excess of loss under the risk measure variance (Problem 1) which can be solved analytically according to [7]. Fig. 1 illustrates that the MOEA fronts have converged well to the true front. When looking at the solution parameters we gain more detailed insights into the trade-off between QS and XL. Similar to the following problem instances (cf. Fig. 3 for instance) we observe a breakage for QS at a certain level, meaning that in riskier solutions no QS protection is applied at all. Additionally we are able to confirm the theorem of Borch/Kahn/Pesonen [5] which states that, given our assumptions, XL is always optimal. These results underpin the capability of MOEAs to converge to the true Pareto front for more complex problems which cannot be solved directly.

Next we extend Problem 1 by allowing the additional use of stop loss contracts (Problem 2). The resulting Pareto front (cf. Fig. 2) is not as smooth any more. The parameters QS and XL exhibit a breakage (cf. Fig. 3 and 4) that has already been observed for QS in Problem 1. SL however represents a rather steady function and is always used (Fig. 4)¹⁰. This leads to:

Observation 1: *In the variance model with a combination of QS, XL and SL, SL is always optimal for the insurer under the expected value principle, while QS and XL are not.*

Moving on to modern risk measures, we first optimize combinations of QS and XL under VaR_α and CVaR_α , respectively (Problem 3 and 4)¹¹. When analyzing the

¹⁰ L shows strong dispersion however where QS and XL are used simultaneously.

¹¹ Due to space restrictions we do not provide figures for these two problems.

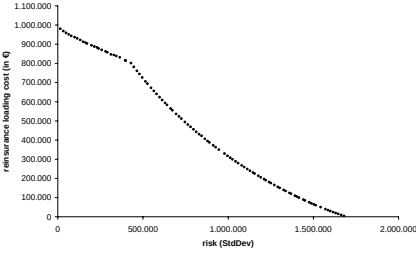


Fig. 2. Optimal fronts for QS, XL and SL

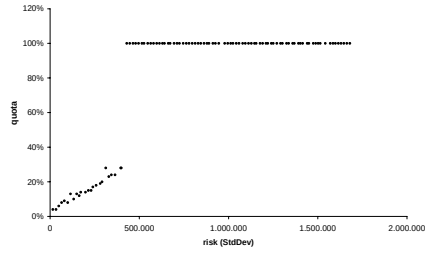


Fig. 3. Quota share a (Problem 2)

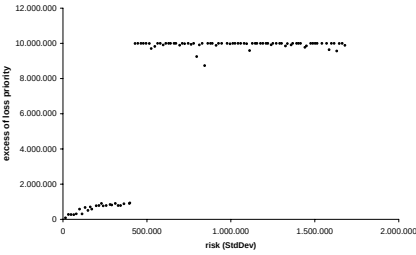


Fig. 4. Excess of loss R (Problem 2)

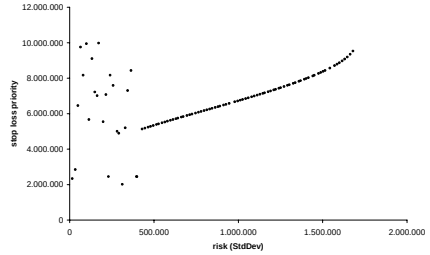


Fig. 5. Stop loss L (Problem 2)

parameters we find the following surprising observation that is different from the above cited theorem of Borch/Kahn/Pesonen in the simple variance model case:

Observation 2: *Under the expected value principle with the risk measures VaR_α or $CVaR_\alpha$ and QS and XL as available reinsurance contracts only combinations of QS and XL are optimal for the insurer.*

Finally we examine the particularly difficult problem (cf. [7]) to combine QS, XL, and SL as in Problem 2 together with VaR_α and $CVaR_\alpha$, respectively (Problem 5 and 6). We present the best Pareto fronts for Problem 5 after a typical run of the two MOEAs in Fig. 6 and the parameter analysis in Fig. 7, 8 and 9. Based on these results we make the following two observations¹²:

Observation 3: *Under VaR_α or $CVaR_\alpha$ with the expected value principle and QS, (unlimited) XL and (unlimited) SL, SL is always optimal for the insurer, while QS and XL are not.*

Based on our assumptions we thus conclude that SL insurance is superior to both QS and XL with QS being slightly superior to XL in certain problem settings. At last it becomes evident that up to the middle of the risk profile all three contract types are used. Therefore we make the concluding observation:

Observation 4: *If the stop loss priority is fixed at a desired risk level VaR_α or $CVaR_\alpha$ and only SL is used, there exist cases where it is advantageous to additionally contract a QS and XL protection for the insurer.*

¹² The results for Problem 6 ($CVaR$) were rather similar and thus are not displayed.

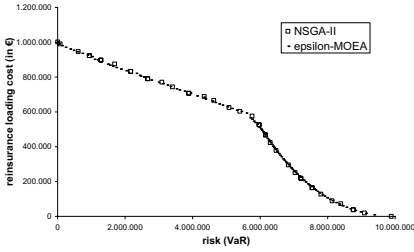


Fig. 6. Optimal for QS, XL, SL

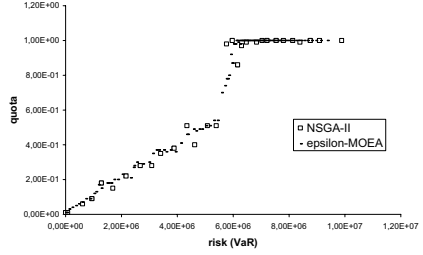


Fig. 7. Quota share a (Problem 4)

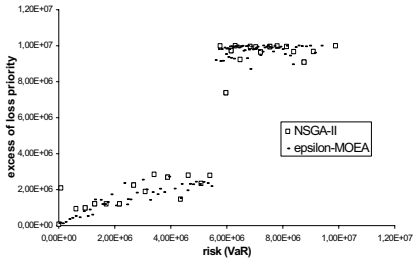


Fig. 8. Excess of loss R (Problem 4)

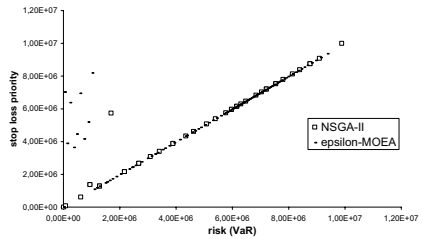


Fig. 9. Stop loss L (Problem 4)

For both MOEA's we use the same metrics spacing and coverage to measure their performance in all problem instances. Spacing values of the NSGA-II are about 0.028 while ε -MOEA yields 0.006. The NSGA-II dominates 0% resp. 4% of the individuals of the ε -MOEA front in the mean, while the latter dominates about 26% of the NSGA-II individuals. Additionally, the ε -MOEA benefits from its growing archive that provides a denser front than the NSGA-II. Hence ε -MOEA provided the results with the highest quality.

5 Conclusion and Outlook

With our multi-objective heuristic approach we are able to enhance existing theoretic findings that have not been solved mathematically yet. Furthermore we derive formerly unknown and highly relevant implications for the practical allocation of reinsurance programs.

Our presented approach is very flexible with respect to real world modifications concerning e.g. risk measurement, cost calculation principles and different combinations of reinsurance treaties. Additionally restrictions through the new capital guidelines for insurers (Solvency II) for instance can easily be integrated into the model setup.

6 Acknowledgements

The authors would like to thank BGV|Badische Versicherungen for providing the data for the computations in this paper. Moreover, the authors would like to thank GILLARDON AG financial software for partial support of their research. Nevertheless, the views expressed in this work reflect the personal opinion of the authors and not necessarily the opinion of BGV|Badische Versicherungen or GILLARDON AG financial software.

For the implementation of the objective functions we partially used code from the GNU Scientific Library. We thank the authors of this library for free distribution of the source code on the WWW. Finally, we are very grateful to Kalyanmoy Deb for providing the source code for the NSGA-II and the ϵ -MOEA on the WWW.

References

1. Carter RL (1995) Reinsurance. Kluwer Publishing, Brentford, Middlesex
2. Deb K, Agrawal S, Pratap A, Meyarivan T (2000) A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimisation: NSGA-II. Technical Report 2000001, KanGAL, Indian Institute of Technology
3. Deb K, Mohan M, Mishra S (2003) A Fast Multi-objective Evolutionary Algorithm for Finding Well-Spread Pareto-Optimal Solutions. Technical Report 2003002, KanGAL, Indian Institute of Technology
4. de Lourdes Centeno M (1985) Some Theoretical Aspects of Combinations of Quota-share and Non-Proportional Reinsurance Treaties. The British Library Document Supply Centre
5. Mack T (1997) Schadenversicherungsmathematik. Verlag Versicherungswirtschaft, Karlsruhe
6. Oesterreicher I, Mitschele A, Schlottmann F, Seese D (2006) Comparison of Multi-Objective Evolutionary Algorithms in Optimizing Combinations of Reinsurance Contracts. In: Proceedings of the 8th annual conference on Genetic and evolutionary computation GECCO '06. ACM Press, New York, USA
7. Verlaak R, Beirlant J (2002) An Optimal Combination of Several Reinsurance Protections on an Heterogeneous Insurance Portfolio. 6th International Congress on Insurance: Mathematics&Economics, IME 2002. Lisbon, Portugal

Optimizing Credit Risk Mitigation Effects of Collaterals Under Basel II

Marc Gürtler, Dirk Heithecker, and Martin Hibbeln

Department of Finance, TU Braunschweig, Germany

`marc.guertler@tu-bs.de`

`d.heithecker@tu-bs.de`

`martin.hibbeln@tu-bs.de`

Summary. One of the main differences in measuring capital requirements for credit risk due to the Internal Ratings Based Approach (IRBA) of the new regulatory Capital Standards (Basel II) is the enlarged acceptance of collaterals in order to reduce (supervisory) credit risk. Whereas in the Advanced IRBA own models for estimating the effects of collaterals can be used, in the Foundation IRBA concrete formulas are given. However, these regulations only explain the case of a single claim collateralized by a single asset or guarantee. If numerous collaterals are available for multiple loans, we receive an allocation problem of collaterals to loans with the objective of keeping regulatory capital low. In this article we model the corresponding optimization problem. Since numerical standard procedures often converge towards the wrong local minimum, we show how to apply evolutionary algorithms as well as simulated annealing.

1 Introduction

With starting of Basel II in 2007/2008 banks face significant changes in measuring credit risk when using the Foundation or Advanced IRBA in order to quantify the minimum (equity) capital needed to meet supervisory requirements [1]. Thus, regulatory credit risk of a loan is defined by the Unexpected Loss (UL) and the Expected Loss (EL), which both depend on the Probability of Default (PD) and the Loss Given Default (LGD). The variable PD quantifies the possibility of a potential default whereas the variable LGD measures the percentage loss in the event of default. The parameters PD and LGD not only depend on the debtor's individual ability to pay his liabilities but also on the debtor's collaterals and guarantees, respectively, to safeguard the bank against potential losses. Different from the Advanced IRBA, in the Foundation IRBA the supervisory board published specific rules how the risk mitigation effect on a single loan has to be measured.

2 Credit Risk Mitigation and Double Recovery

Concretely, under the Foundation IRBA two types of securities exist, collaterals and guarantees. The eligible collaterals are financial collaterals, receivables, real estate, and other physical assets. For each collateral assigned to a loan the bank has to determine its “risk adjusted” value that quantifies the fully collateralized part of the loan. The weighted sum of the reduced *LGD* of this fully secured part and of the *LGD* of the unsecured part defines the so called “effective” *LGD* of the loan. The risk mitigation effect of guarantees follows an approach of “partly” substituting both the risk parameters *PD* and *LGD* of the loan. Firstly, the capital requirements of the fully guaranteed part of the loan depend on the *PD* of the guarantor (substitution approach) or on the *PD* of both the guarantor and the debtor (double-default effect).[2] Secondly, the bank may use the *LGD* of the guarantor or the debtor. Thus, the effect of double recovery is not allowed.¹ Against this background, the collateral that is directly assigned to the loan can not be used to reduce the *LGD* of the guarantor.

3 Formulation of the Allocation Problem

In praxi multiple loans (of one or more debtors) are often secured by multiple collaterals and guarantees, i.e. we have J collaterals $j = \{1, \dots, J\}$ and K guarantees $k = \{1, \dots, K\}$ for I loans $i = \{1, \dots, I\}$. Thus, we have to allocate these $J + K$ securities to the I loans. In order to keep the total capital requirements (*CR*) low, we aim to determine the optimal coefficients x_{ij} and x_{ik} , which assign the allocation of securities to loans in percent. To set up the optimization problem we need (default) risk weights $W_i^{(CR)}$ of a loan and $W_{ik}^{(CR)}$ of a guaranteed loan that are functions of the one-year probability of default PD_i of the debtor and PD_k of the guarantor. If we minimize *CR*, we have to account for (default) risk weights with respect to *UL* and *EL*. Further, we need the adjusted values C_{ij}^* and G_{ik}^* of the securities. We get²

$$W_i^{(CR)} = F_i \cdot SF + PD_i, \quad W_{ik}^{(CR)} = \min \left(PD_k + F_{ik}^{Sub} \cdot SF; F_{ik}^{DD} \cdot SF \right) \quad (1)$$

$$C_{ij}^* = \max \left(0; \frac{C_j}{k_j^h} \cdot \left(1 - H_j^C - H_{ij}^{FX} \right) \cdot H_{ij}^{MM} \right) \text{ and} \quad (2)$$

$$G_{ik}^* = \max \left(0; G_k \cdot \left(1 - H_{ik}^{FX} \right) \cdot H_{ik}^{MM} \right). \quad (3)$$

We receive the minimization problem³

¹ See [1], paragraphs 284 (ii) and 303.

² For further explanation of the variables see Table 1.

³ Again, see Table 1 for the variables. For simplification we set the current value of the exposure equal to *EAD*, that generally fits for typical bank loans. Further, we use $S(\cdot)$ as the indicator function with $S(s) = 1$ if statement s is true.

$$CR = \min_{x_{ij}, x_{ik} \forall i, j, k} \left(\sum_{i=1}^I CR_i \right) \text{ where} \quad (4)$$

$$CR_i = EAD_i \cdot LGD_i \cdot W_i^{(CR)} \quad (5)$$

$$\begin{aligned} & - \sum_{j=1}^J \delta_{ij}^C \cdot x_{ij} \cdot C_{ij}^* \cdot (LGD_i - LGD_j) \cdot W_i^{(CR)} \\ & - \sum_{k=1}^K \delta_{ik}^G \cdot x_{ik} \cdot G_{ik}^* \cdot LGD_i \cdot \left(W_i^{(CR)} - W_{ik}^{(CR)} \right) \\ & - \sum_{k=1}^K \left(1 - \delta_{ik}^G \right) \cdot x_{ik} \cdot G_{ik}^* \cdot \left(LGD_i \cdot W_i^{(CR)} - LGD_k \cdot W_{ik}^{(CR)} \right) \\ & + \sum_{k,j=1}^{K,J} \delta_{ik}^G \cdot \frac{G_{ik}^* \cdot C_{ij}^*}{EAD_i} \cdot x_{ik} \cdot x_{ij} \cdot (LGD_i - LGD_j) \left(W_i^{(CR)} - W_{ik}^{(CR)} \right) \\ & + \sum_{k,j=1}^{K,J} \left(1 - \delta_{ik}^G \right) \cdot \frac{G_{ik}^* \cdot C_{ij}^*}{EAD_i} \cdot x_{ik} \cdot x_{ij} \cdot (LGD_i - LGD_j) \cdot W_i^{(CR)} \end{aligned}$$

with

$$\delta_{ij}^C = S \left(\sum_{j'=1}^J \left(1 - \delta_{j'}^{FR} \right) \cdot x_{ij'} \cdot C_{ij'}^* \cdot \frac{k_{j'}^h}{k_j^m} + \delta_{j'}^{FR} \cdot x_{ij'} \cdot C_{ij'}^* \leq E_i \right) \forall j \quad (6)$$

$$\delta_{ik}^G = S \left(LGD_i + \sum_{j=1}^J \frac{x_{ij} \cdot C_{ij}}{EAD_i} \cdot (LGD_j - LGD_i) \leq LGD_k \right) \forall i, k \quad (7)$$

subject to

$$EAD_i - \sum_{j=1}^J x_{ij} \cdot C_{ij}^* \geq 0, \quad EAD_i - \sum_{k=1}^K x_{ik} \cdot G_{ik}^* \geq 0 \quad \forall i \quad (8)$$

$$\sum_{i=1}^I x_{ij} \leq 1, \quad \sum_{i=1}^I x_{ik} \leq 1 \text{ and } 0 \leq x_{ij} \leq 1, \quad 0 \leq x_{ik} \leq 1 \quad \forall i, j, k. \quad (9)$$

Capital requirement CR_i of each loan (5) consists of five terms. The first term evaluates the regulatory capital needed for the unsecured loan i . The second term measures the risk mitigation effect from collaterals. The next two terms calculate the capital reduction that emerges from guarantees with and without substituting the LGD . With term four and five we satisfy the requirement of non-allowance of the double recovery effect. The additional function (6) decides on substitution of the LGD . Function (7) evaluates whether the value of the collaterals drops under a required minimum level or not. The constraints (8) and (9) assure that each loan i will not be over-collateralized and that the variables x_{ij} and x_{ik} take reasonable values.

Table 1. Definiton of the relevant variables for the allocation problem

Objective Variables			
x_{ij}	allocation of collateral j to loan i in percent	x_{ik}	allocation of guarantee k to loan i in percent
Input Variables			
C_j	actual value of collateral j	G_k	actual value of guarantee k
PD_i	one-year probability of default of debtor (of loan) i	PD_k	one-year probability of default of guarantor (of guarantee) k
Given Parameters, see [1], esp. paragraphs 176, 200, 272, 284ii, 287, 288, 295, 303			
k_j^h	level for full collateralization due to collateral j	SF	Scaling Faktor
k_j^m	minimum collateralization level of collateral j	LGD_i	LGD of the unsecured loan i
H_j^C	haircut for collateral j	LGD_j	LGD due to collateral j of a fully collateralized loan
H_i^{FX}	haircut for currency mismatch between loan i and coll. j or guar. k	LGD_k	LGD due to guarantee k of the fully guaranteed loan i
H_i^{MM}	haircut for maturity mismatch between loan i and collateral j or guarantee k	F_{ik}^{Sub}	default risk weight of the loan i with guarantee k under the substitution approach
F_i	default risk weight of the debtor i	F_{ik}^{DD}	default risk weight of the loan i with guarantee k under the double default approach
Built-in Variables			
δ_{ij}^C	binary variable, is 1 if the minimum collateralization level of collateral j is reached for loan i	δ_{ik}^G	binary variable, is 1 if the LGD of loan i is not substituted with the LGD of guarantee k
$W_i^{(CR)}$	combined EL and UL (default) weight of loan i	$W_{ik}^{(CR)}$	combined EL and UL (default) weight of loan i with guarantee k
C_{ij}^*	risk adjusted value of collateral j assigned to loan i	G_{ik}^*	risk adjusted value of guarantee k assigned to loan i

4 Implementation of Global Optimization Approaches

To solve the nonlinear optimization problem we make use of several optimization approaches. Concretely, we implement the Nelder-Mead algorithm [3] as a local search technique. Further, we make use of simulated annealing [4], a genetic algorithm [5], and an evolution strategy [6] as global search techniques. For all algorithms we adjust the vectors to get an unrestricted problem. After several parameter trials we used the following settings:

Nelder-Mead Algorithm. As a start simplex we initialize an N-polytope with random variables. For the reflection operator we use the multiplier $\gamma = 0.5$, and for the expansion and contraction operators we choose $\beta = 2$ and $\alpha = 1$. For termination, we check the Root Mean Squared Error ($RMSE$) of the actual simplex ($RMSE < 0.1$) and the number of iterations ($I = 250$).

Simulated Annealing. The algorithm starts with 100 random variables and a temperature of 1000°. For the random movement we fix the maximum radius to 10%, the annealing process proceeds with the proportional cooling scheme using $\delta = 0.9$. The algorithm stops when the temperature falls below 0.1°.

Genetic Algorithm. The initialized population exists of 70 random individuals. These individuals are represented as binary strings (10 bit for each collateral or guarantee) and converted using the Gray code representation. We use an elite of 10 individuals and for the selection process we chose the Sigma-Selection criterion with a multiplier for the standard deviation of $\gamma=2$ combined with the Stochastic Universal Sampling Algorithm (SUS). The reproduction is based on a random 4-point-Crossover and bit by bit mutation with probability $p = 0.1\%$. For termination of the algorithm we check the variation of the fitness values and the total number of generations ($g = 500$).

Evolution Strategy. We implement a (30,300)-ES with strategy parameters for correlated mutation. The initial population starts with $\sigma = 0.03$ for the strategy parameters and with stochastic values for the rotation angles α . The mating selection is implemented with identical probabilities for all parents. The decision parameters are handed by discrete combination whereas the strategy parameters are recombined intermediary. For mutation of the mutation-increment we use the logarithmic Gaussian distribution with $\tau_1 = 0.1$ and $\tau_2 = 0.15$ for the global and the local modification parameters and we add Gaussian distributed values (multiplied by $\beta = 0.09$) to the rotation angles. Using these parameters the mutation is executed before the new generation is determined with discriminatory selection. The algorithm terminates when the distance between the best and the worst fitness value falls below $\delta \leq 1$.

5 A Numerical Example

We applied the presented algorithms to minimize the CR for the simple case of allocating one collateral and one guarantee to two loans. Their concrete characteristics are reported in Table 2. All algorithms were implemented in MATLAB[®] and repeated 50 times to assure the results. In Table 3 we report these results for each algorithm, namely the average capital requirement, the maximum error of all trials compared to the global minimum, the number of global optimum hits and finally the average run-time. It is easily seen that the Nelder-Mead algorithm is very fast but can not assure to find the global optimum. The other algorithms are all capable to find this global optimum. The difference in run-time can be explained by a varying accuracy across algorithms. So in this simple case all tested global search techniques show a comparable performance.

As a further step, we also tested more complex mitigation problems. Here we found the simulated annealing and the Evolution Strategy to be more stable than the Genetic Algorithm.

6 Conclusion

We set up the allocation problem for minimizing capital requirement against the background of the acceptance of collaterals under Basel II. We showed exemplary how to solve the optimization problem with evolutionary algorithms and simulated annealing. In the example under consideration all implemented global search techniques were capable to find the global optimum within reasonable time. In future

Table 2. Input data for the simple example

Loan 1	Loan 2	Collateral	Guarantee
<i>PD</i> 10%	<i>PD</i> 10%	Value 150,000 €	Value 150,000 €
<i>LGD</i> 45%	<i>LGD</i> 45%	Maturity 2y	Maturity 4y
<i>EAD</i> 100,000 €	<i>EAD</i> 300,000 €	Haircut 21.21%	<i>PD</i> 1%
Maturity 2y	Maturity 3y		<i>LGD</i> 45%
Sales > 50 m€	Sales > 50 m€		Sales > 50 m€

Table 3. Results for the impemented algorithms

Algorithms	$\emptyset CR$	Max. Error	% Global Optimum Hits	Run-Time
Nelder-Mead	44835 €	19.8%	84%	0.9 sec
Simul. Anneal.	43634 €	0.01%	100%	28.1 sec
Genetic	43634 €	0.001%	100%	69.9 sec
ES	43634 €	0.006%	100%	43.8 sec

research, it should be examined to what extent the algorithms can be accelerated by parallelization or hybridization.

References

- 1. Basel Committee on Banking Supervision (2006) International Convergence of Capital Measurement and Capital Standards. Basel
- 2. Basel Committee on Banking Supervision (2005) The Application of Basel II to Trading Activities and the Treatment of Double Default Effects. Basel
- 3. Nelder JA, Mead R (1965) A Simplex method for function minimization. Computer Journal 7: 308–313
- 4. Kirkpatrick SC, Gelatt D, Vecchi MP (1983) Optimization by Simulated Annealing. Science 220(4598): 671–680
- 5. Holland JH (2001) Adaptation in Natural and Artificial Systems. Cambridge
- 6. Schwefel HP (1995) Evolution and Optimum Seeking. New York

A New Methodology to Derive a Bank's Maturity Structure Using Accounting-Based Time Series Information

Oliver Entrop¹, Christoph Memmel², Marco Wilkens¹, and
Alexander Zeisler¹

¹ Catholic University of Eichstaett-Ingolstadt, Germany
oliver.entrop@ku-eichstaett.de
marco.wilkens@ku-eichstaett.de
alexander.zeisler@ku-eichstaett.de

² Deutsche Bundesbank, Frankfurt, Germany
christoph.memmel@bundesbank.de

This paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank.

1 Introduction

While over the past few years both banking supervisors and researchers have focussed their attention on banks' credit risk, the spotlight is now being turned again on interest rate risk. One reason for this is its character as a kind of systemic risk: there is evidence that a rise in interest rates affects most banks negatively. An historical example of a banking crisis caused by high interest rates is the 'Savings and Loan Crisis' which occurred in the USA during the 1980s.³ Between 1980 and 1988, 563 of the approximately 4,000 savings and loan institutions failed, while further failures were prevented by 333 supervisory mergers. The total costs of the crisis are estimated at USD160 billion. Recently, the Basel Committee on Banking Supervision published principles for the management and supervision of interest rate risk that go far beyond current practice.⁴ However, few data are available concerning banks' interest rate risk exposure.

Whereas the Deutsche Bundesbank has exclusive information for some sub-samples of banks that is based on bank-internal risk management systems, most methods proposed in literature rely on accounting-based information. Accounting-based information on the banks' maturity structure usually contains information on the amount of positions within certain time bands: the total outstanding amount of a given position is distributed among a number of time bands according to initial maturity and/or remaining time to maturity. Approaches proposed in literature use

³ See [4] for a detailed analysis.

⁴ See [1].

one-point-in-time data, typically the most recent report. In order to derive a cash flow structure and its interest rate sensitivity, a certain distribution within the time bands is assumed.⁵ The Economic Value Model (EVM) of the Federal Reserve assumes a concentration in the middle of a time band.⁶ A similar approach is applied by [7] who quantify the interest rate risk of the Indian banking system. In contrast, [2] and [8] assume a uniform distribution. The main shortcoming of these approaches is the omission of further information about the distribution within a time band, which can either be taken from time series information or additional data sources. We present here a methodology to systematically use different accounting-based data sources and time series information to derive the maturity structure of banks' assets and liabilities.⁷ In a simulation approach we compare our new methodology with models proposed in literature and show that time series can contain important information for inferring the banks' maturity structure.

The model presented here was developed within a joint research project of the Deutsche Bundesbank and the Catholic University of Eichstaett-Ingolstadt. An extended model was applied to quantify the interest rate risk exposure of all German banks. For further details and applications see [3].

2 Model

Let time be discrete: $t \in T = \mathbb{N}_0$. In each $t_{beg} \in T$ the bank invests the amount $X^{pos, t_{beg}, t_{end}}$ in a position pos (such as loans on the asset side or deposits on the liability side) that matures in $t_{end} > t_{beg}$. We will refer to these variables as 'business items'. To keep things simple, we omit possible provisions and premature redemptions here. Under this assumption, the outstanding amount in t that is due in $s > t$ is given by

$$OA(pos, t, s) \sum_{i=0}^t X^{pos, i, s}. \quad (1)$$

The knowledge of the outstanding amounts $OA(pos, t, s)$ for each s and pos is the basis for deriving the maturity and cash flow structure of the bank in t . The cash flow structure in turn determines the bank's on-balance-sheet interest rate risk exposure and makes it possible to apply risk measures like duration or value at risk. Unlike the bank itself, external analysts do not have detailed information on the maturity structure. However, at certain dates the bank reveals some information on its maturity structure by reporting the amount of a position broken down into a certain number of time bands. These contain for each pos the sum of all business within specified ranges of initial maturity or remaining time to maturity, respectively. We assume that in each $t_{obs}^{itm} \in T_{obs}^{itm} \subset T$ there are N report items characterized by the business' initial maturity: ITM_t^n with $n \in \{1, \dots, N\}$ reports the amount of pos^n in t with an initial maturity within the time band $t_{lower}^n < t_{end} - t_{beg} \leq t_{upper}^n$. Analogously, in each $t_{obs}^{rtm} \in T_{obs}^{rtm} \subset T$ there are M items characterized by the business' remaining

⁵ See [6] for an advanced approach using detailed information on the banks' assets and liabilities.

⁶ See [5] and [9]. For an evaluation of the model see [11], [10] and [9].

⁷ Accounting-based data can here include both publicly available and non-publicly available data like confidential reports to the banking supervisors.

time to maturity: RTM_t^m with $m \in \{1, \dots, M\}$ reports the amount of pos^m in t with a remaining time to maturity within the time band $t_{lower}^m < t_{end} - t \leq t_{upper}^m$. Since the amount within a certain time band contains the sum of the relevant business items, the former can be expressed as a linear function of the latter. Thus we obtain one function (restriction) for each amount in a time band and each reporting date, leading to a system of linear equations:⁸

$$\sum_{i=0}^{t_{lower}^n} \sum_{j=t_{lower}^n+1}^{t_{upper}^n} X^{pos^n, t-i, t-i+j} + \sum_{i=t_{lower}^n+1}^{t_{upper}^n-1} \sum_{j=i+1}^{t_{upper}^n} X^{pos^n, t-i, t-i+j} = ITM_t^n \quad \forall t \in T_{obs}^{itm}, n \in \{1, \dots, N\} \quad (2)$$

$$\sum_{i=0}^t \sum_{j=t_{lower}^m+1}^{t_{upper}^m} X^{pos^m, i, t+j} = RTM_t^m \quad \forall t \in T_{obs}^{rtm}, m \in \{1, \dots, M\}. \quad (3)$$

All variables are assumed to be nonnegative:

$$X^{pos, t_{beg}, t_{end}} \geq 0 \quad \forall pos, t_{beg}, t_{end}. \quad (4)$$

Equation (2) represents the restrictions due to reports on the initial maturity and (3) those on the remaining time to maturity. The respective right-hand side denotes the reported items (which are given to the analyst) whereas the left-hand side represents the sum of all relevant business items.

The system of equations given by (2), (3) and (4) restricts the values of the business items $X^{pos, t_{beg}, t_{end}}$ and hence via (1) the bank's maturity structure. The system of equations describes the bank over time. To calculate the bank's maturity structure on a certain date t_{ref} ('reference date'), it can be useful to restrict the sample period und thus the included reports. In general, the system of equations obtained is not uniquely determined, i. e. it is not possible to infer a unique maturity structure using accounting-based data. Thus, further assumptions are necessary. These can be formulated in terms of the optimization of a function ('objective function') on the space of solutions.⁹ Hence, the problem is to optimize a function subject to linear restrictions, and the model can be stated as:

$$\begin{array}{ll} \text{optimize} & F(\mathbf{X}) \\ \text{subject to} & (2), (3), (4), \end{array} \quad (5)$$

where F denotes the objective function and $\mathbf{X}$ the set of all business items $X^{pos, t_{beg}, t_{end}}$.

3 Simulation Results

To analyze whether the integration of time series information improves the ability to infer a bank's maturity structure, we here compare the results of the presented approach with an approach using only one-point-in-time data on a sample of synthetic banks. For the latter we assume a uniform distribution within the time bands

⁸ Here we assume that business due in t is not included in the reports in t .

⁹ See Sect. 3 for a possible objective function.

according to [2].

We restrict our sample period to 72 months. For the simulation, we only consider one position (interbank loans), as the remaining positions are modeled analogously. We take the information from the Monthly balance sheets statistics ('Monatliche Bilanzstatistik') and the Data schedule pursuant to the auditor's report ('Sonderdaten-katalog').¹⁰ The Monthly balance sheets statistics contain, on a monthly basis, the assets and liabilities in time bands according to initial maturity. The position interbank loans is broken down into 4 time bands: i) daily maturing, ii) up to 1 year but not daily maturing, iii) more than 1 year and up to 5 years, and iv) more than 5 years. The Data schedule pursuant to the auditor's report contains, on a yearly basis, a breakdown according to the remaining time to maturity. There are 4 time bands available: i) up to 3 months, ii) more than 3 months and up to 1 year, iii) more than 1 year and up to 5 years, and iv) more than 5 years.

The analysis is performed as follows. We first simulate synthetic banks by specifying 'typical' parameter constellations for the variables $X^{pos, t_{beg}, t_{end}}$ and create synthetic reports that the banks would have reported to the Deutsche Bundesbank. We then apply our model as presented in Sect. 2 to derive an implied maturity structure. As an objective function we assume a possibly constant relative distribution of new business on the different maturities over time, specified as follows:

$$\begin{aligned}
 & \text{minimize} && \sum_{t \in T_{obs}} \sum_{i \in \mathbb{N}} d_{t,i}^2 \\
 & \text{subject to} && \\
 & && \frac{X^{pos, t_{ref}, t_{ref}+i}}{\sum_{j \in \mathbb{N}} X^{pos, t_{ref}, t_{ref}+j}} - \frac{X^{pos, t, t+i}}{\sum_{j \in \mathbb{N}} X^{pos, t, t+j}} \\
 & && = d_{t,i} \quad \forall t \in T_{obs}, i \in \mathbb{N},
 \end{aligned} \tag{6}$$

where t_{ref} denotes the reference date at which the maturity structure is to be estimated and T_{obs} the set of all reporting dates: $T_{obs} = T_{obs}^{itm} \cup T_{obs}^{rtm}$. Since the objective function is quadratic, we obtain a quadratic programming problem.

Here we present the results of 4 sample banks. Bank 1a tracks a constant business strategy. In each month it contracts a constant amount of new business that is uniformly (discretely) distributed among all maturities between 1 month and 10 years. Bank 2a, by contrast, contracts only business with an initial maturity of 2, 4, 6, 8, and 10 years. Additionally, the amount of new business increases linearly over time. Banks 1b and 2b are analogous, but contain a stochastic component. Figure 1 shows the results of the 4 sample banks. The solid line gives the true maturity structure, whereas the other two lines represent the estimated ones. The simulation results show that the new approach is able to infer the maturity structure quite well both in a deterministic and a stochastic setting. The approach using only of one-point-in-time data yields a stepwise constant maturity structure. The jumps correspond to the limits of the time bands according to the Data schedule pursuant to the auditor's report. The new methodology, in contrast, yields a smoothed function that fits the true maturity structure quite well, in particular for the banks 1a and 1b that track a constant maturity structure over time.

Thus time series can contain important information on the banks' maturity structure

¹⁰ For further information on these reports, see www.bundesbank.de.

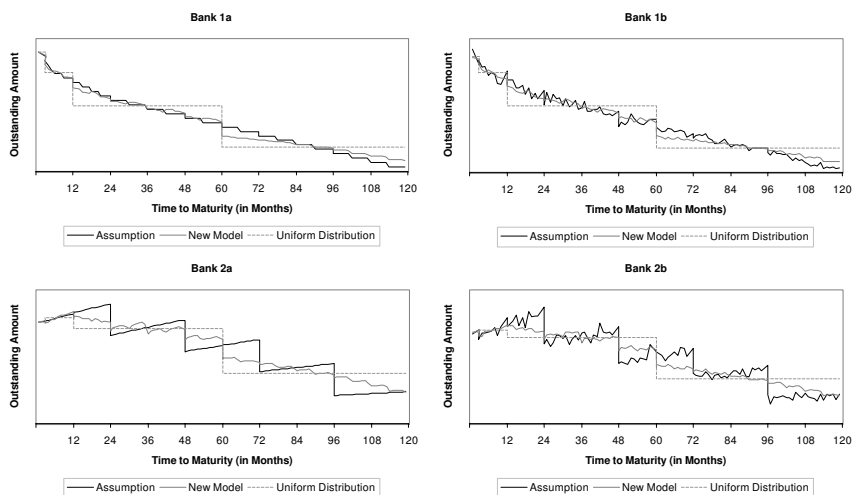


Fig. 1. Simulation results

and should be taken into account when quantifying the interest rate risk exposure of banks using accounting-based data.

4 Concluding Remarks

We presented a methodology to quantify the maturity structure of a bank's assets and liabilities using accounting-based information. In contrast to approaches proposed in literature, the new methodology is able to integrate both the information of different data sources and time series information. In an analysis based on a simulation we showed that the integration of this information yields a better inference of the banks' maturity structure than approaches proposed in literature. In addition to what we presented in this paper, we quantified the interest rate risk exposure of all German banks with an extended model using the data available to the Deutsche Bundesbank.¹¹ We evaluated the results on a sub-sample of banks for which bank-internal data concerning the interest rate risk is available. The empirical results confirm that our approach is able to quantify the interest rate risk exposure of banks more accurately than approaches proposed in literature.

References

1. Basel Committee on Banking Supervision (2004) Principles for the management and supervision of interest rate risk. Basel
2. Bennett D, Lundstrom R, Simonson D (1986) Estimating portfolio net worth values and interest rate risk in savings institutions

¹¹ See [3].

3. Entrop O, Memmel C, Wilkens M, Zeisler A (2006) A linear programming approach to quantify the interest rate risk exposure of German banks. Working Paper, Catholic University of Eichstaett-Ingolstadt
4. Federal Deposit Insurance Corporation (1997) History of the eighties - lessons for the future
5. Houpt J, Embersit J (1991) A method for evaluating interest rate risk in U.S. commercial banks. Federal Reserve Bulletin 77:625–637
6. Office of Thrift Supervision (2000) The OTS net portfolio value model manual
7. Patnaik I, Shah A (2004) Interest rate volatility and risk in Indian banking. IMF Working Paper WP/04/17
8. Planta R. (1989) Controlling interest rate risk. PhD Thesis, Hochschule St. Gallen, Bamberg
9. Sierra G (2004) Can an accounting-based duration model effectively measure interest rate sensitivity? PhD Thesis, Washington University in St. Louis, Saint Louis
10. Sierra G, Yeager T (2004) What does the federal reserve's economic value model tell us about interest rate risk at U.S. community banks? Review/Federal Reserve Bank of St. Louis 86(6):45–60
11. Wright D, Houpt J (1996) The analysis of commercial bank exposure to interest rate risk. Federal Reserve Bulletin February:115–128

Sensitivity of Stock Returns to Changes in the Term Structure of Interest Rates – Evidence from the German Market

Marc-Gregor Czaja and Hendrik Scholz

Catholic University of Eichstaett-Ingolstadt, Germany

`marc.czaja@ku-eichstaett.de`

`hendrik.scholz@ku-eichstaett.de`

1 Introduction

For a long time, interest rates have been considered one of the macroeconomic factors determining stock returns. The role of interest rates in the return generating process of stocks has therefore been extensively investigated in general, but particularly so with regard to financial institutions, which are often deemed to be more sensitive to changes in interest rates than stocks from other industries. Generally, this specific sensitivity has been attributed to i.) the predominant role of financial (i.e. nominal) assets and liabilities on the balance sheets of financial intermediaries and ii.) the maturity transformation performed especially by depository institutions and the resulting maturity mismatch of assets and liabilities (see [14] for an extensive review).

Nevertheless, interest rates also influence the market value of non-financial corporations. First, this is due to the fact that these firms do hold financial assets and especially liabilities as well. Second, interest rates have a decisive impact on investment decisions and thus on future cash flows [1].

This study investigates the stock return sensitivities of both financial and non-financial German corporations to changes in the term structure of interest rates. Most recent contributions pursuing similar research objectives include [11] who investigate a sample of European financial and non-financial corporations, [5] who focus on UK financial institutions, and [6] and [8] for the case of US banks. Studies also investigating a data set similar to ours include the works of [1], [2], and [13] who investigate the sensitivity of German financial and non-financial corporations to changes in interest rates.

The approach most commonly used to estimate the sensitivity of stock returns to changes in interest rates is the two-factor model proposed by [15], which is basically a market model augmented by a second factor intended to capture changes in interest rates. However, [9] show that changes in the term structure are driven by changes in its level, slope, and curvature. Hence, a single interest rate factor will not necessarily capture the entire variability in stock returns caused by changes in the term structure (note the similarity to the duration approach in fixed-income management which has been criticized for the same reason). Therefore, we apply a multi-factor model which

uses changes in level, slope, and curvature of the German term structure of interest rates as factors explaining stock returns.

2 Methodology

[10] propose a parsimonious model of the term structure which describes the yield of a zero bond at time t , $Y_t(T)$, as a function of its time-to-maturity T and four parameters $\beta_{0,t}$, $\beta_{1,t}$, $\beta_{2,t}$, and τ_t .

$$Y_t(T) = \beta_{0,t} + \beta_{1,t} \left(\frac{1 - \exp(-T/\tau_t)}{T/\tau_t} \right) + \beta_{2,t} \left(\frac{1 - \exp(-T/\tau_t)}{T/\tau_t} - \exp(-T/\tau_t) \right) \quad (1)$$

Since the (implicit) loading of $Y_t(T)$ on $\beta_{0,t}$ is one, any change in this factor affects all yields equally and thus causes a level shift. Taking limits of (1), we have $Y_t(\infty) = \beta_{0,t}$ and $Y_t(0) = \beta_{0,t} + \beta_{1,t}$. If we define the difference between the long rate and the short rate as the slope of the term structure, we see that the latter is modelled by $-\beta_{1,t}$. Note that the loading of $Y_t(T)$ on $\beta_{2,t}$ (the second expression in brackets in (1)) converges to zero for both shorter- and longer-term rates. This factor thus governs especially the behavior of medium-term rates and therefore affects the curvature of the term structure [4]. τ_t lacks such an interpretation; it determines the convergence of the exponential terms in (1).

In order to estimate the sensitivity of stock returns to changes in interest rates, we apply the extended factor model proposed by [3] using monthly data. In this model, changes in the parameters of (1) are employed as measures of changes in the shape of the term structure.

$$r_{i,t} = \alpha_i + \beta_{i,L} \Delta L_t + \beta_{i,S} \Delta S_t + \beta_{i,C} \Delta C_t + \beta_{i,M} r_{M,t} + \varepsilon_{i,t} \quad (2)$$

Equation (2) assumes that the returns of stock i in excess of the one-period risk-free interest rate, $r_{i,t}$, are generated by a linear four-factor model. α_i is a constant and $\varepsilon_{i,t}$ is a mean-zero residual. ΔL_t is the first difference of the level factor $\beta_{0,t}$, ΔS_t is the change in the slope factor $\beta_{1,t}$, and ΔC_t captures the change in the curvature factor $\beta_{2,t}$. $r_{M,t}$ is the excess return of a market index. Finally, the β_{iS} are the estimated sensitivities of the i^{th} asset to the respective risk factors.

Correlations between the explanatory variables range from $-.386$ to $+.073$. Despite these relatively low correlations, we apply an orthogonalization procedure to the explanatory variables prior to estimating (2) (for details see [11]). Following [9], we consider the level factor to be the most important interest rate factor, the slope factor to be the second, and the curvature factor to be the least important interest rate factor. Hence we orthogonalize the factors in this order. Finally, to capture market-wide shocks not included in the parameters of the model by [10], we employ a residual market factor where correlations with the interest rate factors have been removed. As a consequence, since the explanatory variables in (2) are linearly independent by construction, we can decompose the stock return variance into the sum of factor variances (times the respective squared sensitivity) and the variance of the error term.

$$\begin{aligned} \text{Var}(r_{i,t}) = & \hat{\beta}_{i,L}^2 \text{Var}(\Delta L_t) + \hat{\beta}_{i,S}^2 \text{Var}(\Delta S_t) + \hat{\beta}_{i,C}^2 \text{Var}(\Delta C_t) + \\ & + \hat{\beta}_{i,M}^2 \text{Var}(r_{M,t}) + \text{Var}(\varepsilon_{i,t}) \end{aligned} \tag{3}$$

Dividing (3) by $\text{Var}(r_{i,t})$ yields the percentage contributions of factors to explaining the total variance of excess stock returns.

3 Data

Our data set covers monthly observations over the period 1974–2002. We use bond price quotes from essentially credit risk-free bills and bonds issued by the Federal Republic of Germany to estimate (1). Technically, we adopt the approach of [4] to estimating (1). For each month t the following three-step procedure is applied. First, we extract “unsmoothed Fama-Bliss” zero bond yields as estimates of $Y_t(T)$ from all bond prices available [7]. Then, using a constant value of 1.5627 for τ_t , we calculate the values of the regressors of (1). Finally, the previously estimated zero bond yields are regressed on the regressors using OLS to estimate $\beta_{0,t}$, $\beta_{1,t}$, and $\beta_{2,t}$.

On average, we estimate a long-rate of approx. 7.4%, a positive slope of 186.6 basis points, and a negative deviation of medium term rates (i.e. curvature) of 101.1 basis points.

Table 1. Descriptive statistics, Nelson-Siegel factors [in %]

	Mean	Std.-Dev.	Min.	Max.
$\hat{\beta}_{0,t}$ (Level)	7.368	1.147	4.870	10.152
$\hat{\beta}_{1,t}$ (Slope $\times -1$)	-1.866	2.016	-6.236	3.762
$\hat{\beta}_{2,t}$ (Curvature)	-1.011	2.453	-6.002	7.425

Monthly total return data for the Datastream German industry indices listed in Table 2 has been collected for the period 1974–2002. Datastream’s broad “Germany Total Market” Index is employed as the market factor. All time series of returns are adjusted for the one-month German interbank rate.

Table 2. Descriptive statistics, stock market indices (excess returns), 1974–2002

	Mean [%/month]	Std.-Dev. [%/month]	Correl. w./ mkt. factor	Avg. share of tot. mkt. cap.
Automobiles	0.409	7.911	0.748	0.097
Chemicals	0.520	5.982	0.789	0.126
Consumer Goods	0.191	5.824	0.753	0.037
Health Care	0.320	4.538	0.731	0.046
Utilities	0.395	3.523	0.479	0.042
Tot. Non-Financials	0.303	5.120	0.974	0.713
Banks	0.288	6.422	0.851	0.148
Insurances	0.666	7.247	0.831	0.119
Total Financials	0.423	6.276	0.900	0.287
Total Market	0.321	5.212	1.000	1.000

4 Empirical Results

We estimate the sensitivity of the industry indices listed in Table 2 to changes in the shape of the term structure using the orthogonalized time series of the factors in (1) and a residual market factor as explanatory variables. The coefficients and robust t-statistics of (2) are shown in Table 3, while the contribution of these factors to explaining the total variance of the respective indices (3) are given in Table 4.¹

From Table 3, we find the vast majority of industries to have a negative exposure to changes in the shape of the term structure. This is consistent with economic theory, since any positive change in the parameters of the model by [10] will lead to an increase in interest rates and thus to an increase in the cost of capital, assuming all else is equal. While the exposure to level changes is significant for all industries, we find significant exposures to the slope factor only in the case of insurance companies and the utility sector. Changes in the curvature significantly affect financial institutions in general, the utility sector, and the consumer goods industry as well.

Table 3. Sensitivities to changes in the shape of the term structure, 1974–2002

	$\hat{\alpha}_i$		$\hat{\beta}_{i,L}$		$\hat{\beta}_{i,S}$		$\hat{\beta}_{i,C}$		$\hat{\beta}_{i,M}$	
Automobiles	0.003		-4.530	***	-1.290		-0.031		1.142	***
Chemicals	0.005	**	-4.525	***	0.114		-0.079		0.904	***
Consumer Goods	0.001		-3.290	***	-0.084		-0.489	*	0.848	***
Health Care	0.003		-2.677	***	0.412		-0.031		0.641	***
Utilities	0.004	**	-3.818	***	-0.610	*	-0.724	***	0.296	***
Tot. Non-Financials	0.003	***	-4.062	***	-0.327	**	0.093		0.963	***
Banks	0.002		-6.324	***	-0.184		-0.407	**	1.027	***
Insurances	0.006	***	-5.804	***	0.948	**	-0.882	***	1.147	***
Total Financials	0.003	**	-6.034	***	0.255		-0.661	***	1.067	***

Note: */**/** denotes statistical significance at the 10%/5%/1% level based on Newey-West robust standard errors.

In terms of the percentage of variance explained, we find that changes in the term structure explain the highest percentage of total variance for the utility sector. This is consistent with prior research by, e.g., [12], who attributes the high relevance of interest rates for utility stocks to i.) rate regulations (having effects on future cash flows resembling nominal assets) and ii.) the high debt levels traditionally carried in this industry. While the level factor clearly dominates all other interest rate factors in explaining total variance, at least for utility stocks and the insurance sector, the curvature factor must not be neglected since it explains roughly 2.8% and 1% of total asset return variance, respectively.

¹ In order to conserve space, we only report results of the total sample period 1974–2002. However, because of this relatively long period, it might be possible that structural changes affect the results. [3] provide an investigation of a similar data set but for shorter sample periods as well.

Table 4. Contribution of factors to explaining total variance [in %], 1974–2002

	Level	Slope	Curv.	Σ Term Struc.	Res.- Mkt.	Tot. R^2
Automobiles	2.263	0.552	0.001	2.816	53.600	56.417
Chemicals	3.934	0.007	0.011	3.952	58.476	62.428
Consumer Goods	2.196	0.004	0.460	2.661	54.385	57.046
Health Care	2.392	0.170	0.003	2.566	51.174	53.739
Utilities	8.144	0.625	2.778	11.548	18.282	29.830
Tot. Non-Financials	4.333	0.085	0.022	4.440	90.673	95.113
Banks	6.720	0.017	0.264	7.001	65.983	72.984
Insurances	4.432	0.356	0.972	5.760	64.451	70.212
Total Financials	6.408	0.034	0.730	7.172	74.693	81.865

5 Summary

We investigate here the sensitivity of German corporations to changes in interest rates. Contrary to previous studies, we explicitly address the exposure to those factors governing changes in the shape of the term structure, namely to changes in level, slope, and curvature. Consistent with prior research by, e.g., [11], we find evidence of interest rate risk being generally of higher relevance for financial institutions compared to non-financial corporations. However, the utility sector yields a different picture. This industry is the most affected by changes in interest rates in our sample. Nevertheless, all industries considered display a significant exposure to changes in the level of the term structure, which explains why this factor is found to be the most important interest rate factor explaining the variability in stock returns. Still, depending on the industry, the curvature – and to a lesser degree the slope factor – also contribute to explaining the return variance of German stocks.

References

1. Bartram SM (2002) The Interest Rate Exposure of Nonfinancial Corporations. *European Finance Review* 6:101–125
2. Bessler W, Opfer H (2003) Empirische Untersuchung zur Bedeutung makroökonomischer Faktoren für Aktienrenditen am deutschen Kapitalmarkt. *Financial Markets and Portfolio Management* 17:412–436
3. Czaja MG, Scholz H, Wilkens M (2006) Interest Rate Risk of German Financial Institutions – The Impact of Level, Slope, and Curvature of the Term Structure. Working Paper, Catholic University of Eichstaett-Ingolstadt
4. Diebold FX, Li C (2006) Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics* 130 2:337–364
5. Dinienis E, Staikouras SK (1998) Interest Rate Changes and Common Stock Returns of Financial Institutions: Evidence from the UK. *European Journal of Finance* 4:113–127
6. Elyasiani E, Mansur I (1998) Sensitivity of the Bank Stock Returns Distribution to Changes in the Level and Volatility of Interest Rate: A GARCH-M Model. *Journal of Banking and Finance* 22:535–563
7. Fama EF, Bliss RR (1987) The Information in Long-Maturity Forward Rates. *American Economic Review* 77:680–692

8. Fraser DR, Madura J, Weigand RA (2002) Sources of Bank Interest Rate Risk. *The Financial Review* 37:351–368
9. Litterman R, Scheinkman J (1991) Common Factors Affecting Bond Returns. *Journal of Fixed Income* 1991 6:54–61
10. Nelson CR, Siegel AF (1987) Parsimonious Modeling of Yield Curves. *Journal of Business* 60:473–489
11. Oertmann P, Rendu C, Zimmermann H (2000) Interest Rate Risk of European Financial Corporations. *European Financial Management* 6:459–478
12. O’Neal ES (1998) Why Electric Utility Stocks are Sensitive to Interest Rates. *The Financial Review* 33:147–162
13. Scholz H, Simon S, Wilkens M (2006) Die Zinssensitivität börsennotierter deutscher Finanzdienstleister – Eine empirische Untersuchung unter Berücksichtigung alternativer Zinsfaktoren. Forthcoming in: *Kredit und Kapital*.
14. Staikouras SK (2003) The Interest Rate Risk Exposure of Financial Intermediaries: A Review of the Theory and Empirical Evidence. *Financial Institutions, Markets & Instruments* 12:257–289
15. Stone BK (1974) Systematic Interest Rate Risk in a Two-Index Model of Returns. *Journal of Financial and Quantitative Analysis* 9:709–72

The Valuation of Localization Investments with Real Options: A Case from Turkish Automotive Industry

Gül Gökay Emel and Pinar Özkeserli

Department of Business Administration, Uludag University, Bursa, Turkey
ggokay@uludag.edu.tr
pinar.ozkeserli@gmail.com

Summary. Localization investments are made mainly to increase the domestic content of manufacturing, thereby to reduce the costs and import dependency. Since localization investments include technology transfer, they positively affect the developing economies. The arrival of the new technology to the country has a positive impact on development of the equipment, workforce and the final product. It is often applied in automotive industry. Through localization investments not only improvement of OEMs, but also development of involved subsidiary industry can be induced. As a developing economy itself, Turkey encourages these kinds of investments and gives governmental incentives. The problem here is the valuation of these investment projects accurately by including all strategic impacts. Just because localization projects involve many future growth opportunities, conventional valuation tools cannot value those embedded opportunities correctly and the value of the project mostly appears negative or very low. This situation can cause a misjudgment at the incentives stage. This paper analyzes the localization project, first with conventional NPV analysis, then with real options analysis and finally compares the results. As for the options analysis, the results of potential involvement of various option types such as sequential options approach are used in the valuation of a case from Turkish automotive industry.

1 Introduction

Localization investments are mainly understood as import substitutions which minimize the foreign exchange loss, the burden of supply processes and foreign dependency. Frequently it is used to bring high technology into the country and help local industries to create their own technology and innovation basis. Also, they open new markets to local manufacturers and create chain reactions. They not only help the OEM's to improve but also help even the second and third degree tiers to grow. Here "growth" includes the R&D departments, the equipments and the workforce. Strategic importance of localization investments is more significant than their financial revenues. In a dynamic and uncertain business environment there is a need

for strategic valuation methods in decision-making rather than static methods. Real options is a methodology that includes dynamism[6]. Real options analysis helps to value the embedded opportunities in the investment project, which may not be valued with the conventional valuation techniques. Real options analysis helps to take the advantage of upside potentials and to avoid downside risks [3]. The lack of capturing these strategic opportunities causes misevaluation of the projects and sometimes even stops the improvement of the developing economies.

Turkey has a promising position in the automotive industry, mainly within the East Europe/Middle East North Africa (MENA) region [10]. Turkey also has low localization ratios then its competitors in the East Europe and a very high import dependency mostly on the crucial automotive vehicle parts. Turkey should start to build its own innovation culture and creativity in order to gain a competitive position in global economies. R&D and many other manufacturing processes must be localized and costs must descend with the advantages of producing and transporting locally. This is also accepted and supported by the Turkish government. In this sense the investment valuations should be done precisely in order to contribute to the Turkish economy and effective usage of the state budget resources.

This paper handles the localization investments as a strategic decision making process which can not be valued with conventional valuation tools but with real options. In this study, a case from the Turkish automotive industry is solved first with conventional NPV analysis, then with real options and a comparison has been made.

2 Classic Valuation Methods

Discounted cash flow (DCF) analysis evaluates the present value of any stream of future cash flows to a specific time and allows comparing two streams of cash flows in terms of their financial value [4]. It's been recognized that standard DCF criteria often undervalued investment opportunities, leading to myopic decisions, underinvestment and eventual loss of competitive position. In particular, DCF methods ignore the operating flexibility that gives management the option to revise decisions while a project is underway which often lead to short term results and ignore projects' longer term prospects [9].

The value of active management and project interactions is not captured in the standard framework of an accept-or-reject decision and post-audit analysis. In a turbulent business environment, new opportunities and old plans must be revised or abandoned. Option pricing combined with dynamic strategic planning and control models are better to capture these sources of value. The presence of managerial flexibility can make the investment opportunity economically attractive. Contrary to conventional popular belief, higher uncertainty, greater interest rates, or more time before undertaking a project does not necessarily make an investment opportunity less valuable [8].

3 Real Options and Sequential Compound Options

Sustainable competitive advantages empower companies with valuable options to grow through future profitable investments and to respond to unexpected adversities

or opportunities in a changing business environment [8]. Real option analysis may help to explain the importance of managerial flexibility and future opportunities to decision makers. A real options approach incorporates a learning model such that management makes better and more informed strategic decisions when some levels of uncertainty are resolved through the passage of time [6]. Uncertainty widens the range of potential outcomes. Contingent decision making limits the loss of a bad outcome, so that greater uncertainty has only one sided affect of increasing upside potential [1].

Researchers have suggested that any contingent claim on an asset, whether traded or not, can be priced in a world with systematic risk by replacing its actual growth rate with a certainty-equivalent rate and the behaving as if the world were risk-neutral. Option pricing requires six basic input parameters: current value of the underlying asset, the rate of forgone earnings, the variance of the rate of return on the underlying asset, the cost of pursuing the opportunity under consideration, the risk free interest rate and the time remaining until the opportunity disappears [9].

At the strategic level, many projects are not independent. They are better seen as links in a chain of interrelated projects, the earlier of which are prerequisites for the ones to follow. Project interdependency may have considerable strategic importance, since it may justify acceptance of projects with negative static NPV on the basis of their potential to open up new investment opportunities in the future. Options whose pay offs are other options are called compound options. A sequential compound option means that the execution and value of future strategic options depend on previous options in sequence of execution. Compound options may be undertaken not only for their directly measurable cash flows but also for the new opportunities they may open up. In general, such inter-project options are associated with investments intended to place the firm on a growth path or to improve its strategic position in the industry. Integrating the real-options approach with a strategic-investment mix model may make it possible to quantify the various strategic sources of value embedded in collections of projects [8]. A sequential compound option exists when a project has multiple phases and latter phases depend on the success of previous phase [6]. The reason for using this kind of option in this paper is dependency of the latter phases of the investment heavily on the approval from the headquarters of the OEM's.

4 A Localization Investment Case from Turkey

4.1 The Definition

The decision to make or buy is arguably the most fundamental component of manufacturing strategy [7]. In this case, an OEM settled in Bursa, Turkey has a localization investment project for developing a technology for dying of metal parts for long corrosion proof life to make the buy-or-make decision. The OEM needs a governmental incentive. The project has three phases and includes the R&D of a new dying technique, the set-up investment and an expansion investment anytime after the fourth year. The first phase includes the initial expenditure for the R&D development, costs 1.000.000 TRY (1 TRY $\approx$ 0,54 EUR) . The second phase includes the set-up investment at the first year and costs 3.000.000 TRY. To go through

the second phase, the first phase which is the process of dying needs to pass the non-corrosiveness test and be approved by the headquarters of the OEM. The set up of the plant is easy to implement, for this reason set-up duration is negligible. The third phase includes the expansion opportunity anytime after the fourth year assumed to provide 70% percent extra capacity, costs 1.500.000 TRY. The risk free rate is 17%, the volatility of the cash flows is 9% [11] and the life of the project is 5 years . The discount rate for the NPV analysis is 22%.

4.2 The Valuation and Result

Firstly, the conventional NPV technique is used in the valuation. Valuation of the investment is found to be -41.082TRY, which would cause the investment to be rejected.

$$NPV = -1.000.000 + \frac{-1.600.792}{(1 + 0,22)^1} + \dots + \frac{2.203.059}{(1 + 0,22)^5} = -41.082 \text{ TRY}$$

The gross present value of the expected cash inflows by a static valuation is found to be 2.892.902 TRY. The same valuation is made with real option analysis. This is the underlying value of the option. Table 1. shows the evaluation lattice.

Table 1. Lattice Evolution of the Underlying

	1	2	3	4	5
2893,0	3165,4	3463,5	3789,7	4146,6	4537,1
	2644,0	2893,0	3165,4	3463,5	3789,7
		2416,4	2644,0	2893,0	3165,4
			2208,5	2416,4	2644,0
				2018,4	2208,5
					1844,7

Because the option is sequential the latter phases are valued first, consequent to the backward induction method. Table 2. and Table 3. show the options valuation steps. Table 4. shows the combined option analysis from Table 2. and Table 3.

As seen from the lattices, the result of the analysis is 630.040 TRY while the valuation of conventional analysis’s result was a negative 41.082 TRY. It is appropriate to make the investment due to the results from the real option analysis. The OEM can make the investment in order to start manufacturing rather than importing.

5 Conclusions

Dependency on technology, inability of creating innovation and new technology and unstable exchange rates usually has negative impacts on unit costs. Investment systems involve many options. The lack of appropriate valuation techniques to value the embedded opportunities in projects can affect the vital investments by ignoring the strategic benefits especially in developing economies. The localization target set for the OEM’s in developing countries is considered to be 60%. Based on this ratio,

Table 2. Option Valuation Step – Expand After 4th Year

	1	2	3	4	5
4161,06	4480,5	4820,4	5177,0	5549,3	6213,1
				<i>expand</i>	<i>expand</i>
	3587,3	3850,4	4115,8	4388,0	4942,5
				<i>expand</i>	<i>expand</i>
		3056,1	3229,3	3418,1	3881,3
				<i>expand</i>	<i>expand</i>
			2451,7	2607,9	2994,8
				<i>expand</i>	<i>expand</i>
				2018,4	2254,4
				<i>keep open</i>	<i>expand</i>
					1844,7
					<i>keep open</i>

Table 3. Option Valuation Step – Invest 1st Year – Invest for R&D

630,04	1480,5
	587,3

Table 4. Combined Option Valuation Lattice

	1	2	3	4	5
630,04	1480,5	4820,4	5177,0	5549,3	6213,1
				<i>expand</i>	<i>expand</i>
	587,3	3850,4	4115,8	4388,0	4942,5
				<i>expand</i>	<i>expand</i>
		3056,1	3229,3	3418,1	3881,3
				<i>expand</i>	<i>expand</i>
			2451,7	2607,9	2994,8
				<i>expand</i>	<i>expand</i>
				2018,4	2254,4
				<i>keep open</i>	<i>expand</i>
					1844,7
					<i>keep open</i>

9 billion\$ production value, 5.4 billion\$ added value, 3 billion\$ export income and 450 million\$ investment potential is expected in these countries [2]. However, the localization ratio in Turkey is less than the average when 1997 data of 49% and 2002 data of 22% are considered [5]. Inevitably, the implementation of these kinds of investments is of vital importance for Turkey for effective use of limited resources.

In this paper a localization investment in Turkish automobile industry is analyzed first with NPV method and then with real options method and a comparison has been made. Different results obtained from two different analysis show that, real options valuation can help decision makers to see the embedded opportunities in strategic investments and prevent the rejection of future opportunities in economically vital projects. Real options have the potential to make a significant difference in the area of competition and strategy.

References

1. Amran Martha, Kulatilaka Nalin (1999) Real Options: Managing Strategic Investment In an Uncertain World. Harvard Business School Press.
2. Devlet Planlama Teşkilatı, Dokuzuncu Kalkınma Planı (2007-2013) Otomotiv Sanayii Özel İhtisas Komisyonu Raporu.
3. Ford David N., Sobek Durward K. (2005) Adapting Real Options to New Product Development by Modeling the Second Toyota Paradox. IEEE Transactions on Engineering Management 52: 175-185.
4. Howell Sydney (2001) Real Options: Evaluating Corporate Investment In a Dynamic World. Financial Times Prentice Hall, GB.
5. <http://vizyon2023.tubitak.gov.tr/teknolojiongorusu/paneller/makinevemalzeme/raporlar/Ek6a.pdf>
6. Mun Jonathan (2002) Real Options Analysis: Tools and Techniques For Valuing Strategic Investments and Decisions. John Wiley & Sons.
7. Shtub Avraham, Bard Jonathan F., Globerson Shlomo (1994) Project Management Engineering, Technology and Implementation. Prentice Hall.
8. Trigeorgis Lenos (1995) Real Options In Capital Investment: Models, Strategies, and Applications. Praeger.
9. Trigeorgis Lenos (1996) Real Options: Managerial Flexibility and Strategy In Resource Allocation. MIT Press.
10. UNIDO Industrial Development Report (2005) Vienna.
11. Automotive Manufacturers Association, www.osd.org.tr.

Health and Life Science

ILP Models for a Nurse Scheduling Problem

Bettina Klinz¹, Ulrich Pferschy², and Joachim Schauer¹

¹ Institut für Optimierung und Diskrete Mathematik, TU Graz, Austria
`klinz@opt.math.tu-graz.ac.at`

² Institut für Statistik und Operations Research, Universität Graz, Austria
`pferschy@uni-graz.at`

1 Introduction

The widely deplored shortage of qualified personal in the nursing sector all over Europe can be traced to a number of reasons among which the relatively short duration that personal remains on the job ranks prominently. Besides the high psychic pressure involved with many nursing jobs, the degree of job satisfaction is also decreased by irregular working hours and inflexible working schedules.

Many different types of nurse scheduling problems have been treated in the literature (see the excellent literature review by Burke et al. [1]). What basically all these versions have in common is that working shifts have to be assigned to nurses for a certain planning horizon (usually one month), such that the given demand and a large number of legal, departmental and organizational constraints are fulfilled.

The special type of nurse scheduling problem we deal with in this contribution arises in the university hospital of Graz (LKH Klinikum Graz), an institution with a nursing staff of more than 3000. Schedules are currently set up manually by the head nurses of each ward where a ward comprises 10 to 50 nurses. Our main objective is not minimization of costs, but the generation of schedules which are perceived as fair and satisfactorily with regard to general and individual preferences.

The special features making this problem worth to be studied are the following. First of all, the list of available working shifts does not consist only of the traditional early, late and night shift, but of a large variety of overlapping shifts with different starting times and durations. Secondly, the daily demand profile splits each day in a number of demand periods, each with a given minimum number of nurses. The demand periods do not correspond, however, to the time partition caused by the available shifts. Thirdly, there is a significant proportion of part-timers including not only half-time employees, but also other fractions between $1/2$ and 1 . An interesting structural property is the preference for *blocks of duties*, which means that blocks of three or four working days followed by several days off are strongly preferred against single working days alternating with single days off. Additional difficulties result from the fact that individual preferences of the nurses should be taken into account and from the fact that in some planning periods no feasible solution might exist due to too many nurses on holidays or on sick leaves.

In this extended abstract we will describe the basic features of two mathematical models for the version of the nurse scheduling problem that showed up in our application. Further details will be presented in the full paper version.

2 Problem Description

A full description of all constraints that have to be taken into consideration when designing a feasible monthly schedule in our application is beyond the scope of this paper. Hence, we concentrate on the most important structural properties a schedule has to fulfill and simplify or even ignore some of the less important ones.

In our case, a set of n nurses must be scheduled for a 30 day planning horizon. The nurses are partitioned into four sets corresponding to the type of their contract, namely $1/2$, $2/3$, $3/4$ and full employees.

Working time is assigned by shifts. For each available shift type the time period it covers is given where some shifts contain a break in between. Moreover, for each shift type it is known for which group of nurses it is available. E.g. a shift with 8 hours duration may only be assigned to a full-time or a $3/4$ -employed nurse, whereas a four hour shift may not be assigned to a full-time nurse. Night shifts always last for 12 hours and can be assigned to any type of nurse.

Demand is given by a partitioning of the day into periods in each of which a certain number of nurses is required to work. Unfortunately, the start and end time of shifts do not coincide with the endpoints of the demand periods which makes overstaffing unavoidable. A real-world example of shift types and demand data is given in Table 1.

Table 1. Shift and demand data of a hospital ward.

shift	start	end	break	hours	demand period	number of nurses	
F24	06:30	14:30	-	8	06:30-08:00	weekdays	weekends
F20	06:30	12:30	-	6	08:00-11:30	4	3
H18	09:00	18:00	12:00-13:00	8	11:30-12:00	5	4
H33	06:30	18:00	10:30-14:00	8	12:00-14:30	7	6
S14	11:00	19:00	-	8	14:30-18:45	6	5
S38	13:00	19:00	-	6	18:45-06:30	2	2
N8	18:45	06:45	-	12		2	1

The following constraints are *hard constraints* that have to be fulfilled by any feasible schedule.

1. Every third weekend must be free.
2. The total working time may deviate from the required working time RWT (e.g., number of working days per month times eight hours for a full-time nurse) by at most 20 hours.
3. The working time per week is at most 60 hours.
4. Between two shifts of work there must be at least 11 hours of free time.
5. After working a sequence of night shifts, at least 2 free days must follow.

6. At most five days in a row may be working days.

The next four constraints are *soft constraints* that are desirable, but which are not strictly enforced. The amount of their violation contributes to the perceived quality of a schedule.

1. Every second weekend should be free.
2. Every week there should be a contiguous interval of 36 hours free time including the 24 hours of one day.
3. At most three night shifts may be performed in a row.
4. Weekends should be assigned as one task, i.e., no nurse should work only on Saturday or only on Sunday of a specific weekend.

While the individual preferences regarding a good schedule vary considerably between nurses, it turned out that the blocking structure of working days plays an important role for all nurses: It is highly preferred to work several days in a row followed by several days off instead of having isolated working days. A special block pattern applies for night shifts.

A very important aspect for the acceptance of an automatic planning system is the possibility to include individual wishes. It turned out that the most important wish is to have prespecified days off. Most other wishes can be handled simply by direct exchanges of shifts after the automatic planning has been completed.

3 An Assignment ILP Model

It is commonly accepted that ILP models which cover all aspects of a complex nurse scheduling problem are intractable for current ILP solvers (cf. [1]). Hence, simplified ILP models are required which concentrate on the main structural aspects of the problem and incorporate other features by heuristics and postprocessing. In this paper we will first give a straightforward assignment type formulation which is based on shift and demand periods, but neglects the block structure (the latter will be handled in the model in Section 4).

We are given the available shifts $H_1, \dots, H_S$ with durations $\text{len}(H_k)$ and the demand values d_t for every period t . Let $\{k \mid t \subseteq H_k\}$ be the index set of all shifts containing period t . By introducing the following decision variables

$$x_{ijk} = \begin{cases} 1 & \text{if nurse } i \text{ works shift } H_k \text{ on day } j, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

we easily arrive at the following ILP model.

$$\min \sum_{i=1}^n \sum_{j=1}^{30} \sum_{k=1}^S x_{ijk} \quad (2)$$

$$\text{s.t.} \quad \sum_{k=1}^S x_{ijk} \leq 1, \quad \forall i, \forall j, \quad (3)$$

$$\sum_{i=1}^n \sum_{k:t \subseteq H_k} x_{ijk} \geq d_t, \quad \forall t, \forall j, \quad (4)$$

$$\text{RWT}_i - 20 \leq \sum_{j=1}^{30} \sum_{k=1}^S \text{len}(H_k) x_{ijk} \leq \text{RWT}_i + 20, \quad \forall i, \quad (5)$$

$$\sum_{\ell=j}^{j+5} \sum_{k=1}^S x_{i\ell k} \leq 5, \quad \forall j, \forall i, \quad (6)$$

$$x_{ijk} \in \{0,1\}, \quad \forall i, j, k. \quad (7)$$

The objective (2) minimizes the total number of worked shifts. Alternatively, we may minimize the total working time by multiplying x_{ijk} in (2) by $\text{len}(H_k)$. (3) guarantees that every nurse works at most one shift per day. Demand is enforced by (4) for every time period on every day by summing up over all shifts that fully cover this period. Constraint (5) bounds the total working time of a month, where RWT_i denotes the required working time for nurse i . Finally, (6) makes sure that at most 5 days are worked in a row. Night shifts are not explicitly included in the model, but are contained in the list of shifts.

The desired block structure cannot be easily incorporated in this model. It is, however, possible to discourage isolated working days by introducing additional binary variables and penalizing them in the objective function.

4 A Multi-Commodity Flow Formulation

In order to be able to achieve schedules with a nice block structure, we propose a multi-commodity flow based ILP model which is motivated to some extent by the work of Cappanera and Gallo [2] on airline crew scheduling. We set up an acyclic directed graph $G = (N, A)$ with nodes

$$w_j^\ell, \quad \ell = 1, \dots, 5, j = 1, \dots, 30 - \ell + 1$$

to represent a block of ℓ working days starting on day j and nodes

$$f_j^\ell, \quad \ell = 1, \dots, F, j = 1, \dots, 30 - \ell + 1$$

to represent a block of ℓ free days starting on day j , where F is the maximal number of days without work. Night shifts are modelled by special work nodes corresponding to the pattern of night shifts required in the particular ward.

An arc connects two nodes corresponding to a feasible sequence of a work block and a free block and vice versa, e.g., the arc (w_6^4, f_{10}^2) means days 6,7,8,9 are working days, and days 10 and 11 are free days. The cost of an arc c_a indicates the “disutility” of such a combination. This allows to deal with wishes like having a longer break

after many working days and having a not too short break before a long block of work. While the block structure is inherent in this model, different shifts and modes of employment are dealt with by pre- and postprocessing and tailor-made additional constraints. In particular, for every day j a minimal number of nurses d_j is computed which is required to allow a feasible schedule for this day. The actual assignment of shifts takes place in a heuristic post-processing step.

For every nurse $i = 1, \dots, n$ we define a commodity i with source s^i , sink t^i and a demand of 1. The flow variable $x_a^i \in \{0,1\}$ denotes the flow of commodity i on arc a . Every possible monthly schedule of nurse i consists of the duties corresponding to the nodes on a path from s^i to t^i defined by those arcs a where $x_a^i = 1$.

$$\min z = \sum_{i=1}^n \sum_{a \in A} c_a x_a^i \quad (8)$$

$$\text{s.t.} \quad \sum_{a \in (s^i, \cdot)} x_a^i = 1, \quad \forall i, \quad (9)$$

$$\sum_{a \in (\cdot, t^i)} x_a^i = 1, \quad \forall i, \quad (10)$$

$$\sum_{a \in (\cdot, v)} x_a^i = \sum_{a \in (v, \cdot)} x_a^i, \quad \forall v \in N \setminus \{s^i, t^i\}, \forall i, \quad (11)$$

$$\sum_{i=1}^n \sum_{\ell=1}^5 \sum_{s=j-\ell+1}^j \sum_{a \in (\cdot, w_s^\ell)} x_a^i \geq d_j, \quad \forall j, \quad (12)$$

$$x_a^i \in \{0,1\}, \quad \forall a, \forall i. \quad (13)$$

The objective (8) minimizes the total “unhappiness” over all nurses. (9), (10) and (11) guarantee together that for commodity i one unit of flow is sent along a path from s^i to t^i . (12) makes sure that the total number of nurses working on day j meets the demand d_j . Most individual wishes can be easily modelled by an appropriate choice of the arc costs and flow variables.

The above model can be used to assign a schedule to every nurse such that the daily demand is fulfilled. There is no way, however, to check whether the work load distribution is balanced. It could happen that all working days are assigned to the same d_j nurses and all free time to the other $n - d_j$ nurses. To achieve a certain amount of fairness, an upper bound for the total cost of the flow of each nurse is introduced by adding the constraint

$$\sum_{a \in A} c_a x_a^i \leq \frac{z^{LP}}{n} + C, \quad \forall i, \quad (14)$$

where z^{LP} is the optimal value of the LP-relaxation of (8)–(13) and C an appropriately chosen small constant. Other restrictions may be added to improve the fairness of the generated schedule. Generally, they can be seen as *resource constraints* which limit the amount of inconvenience caused by a certain type of hardship. In our tests we added e.g. constraints to limit the number of night shifts performed by every nurse. As can be expected, the running time to solve the model increases considerably with the number of such resource constraints. However, this increase can be limited by enforcing a suitable ordering of the variables in the branching process of

the ILP solver, namely, branching first on variables which are included in a resource constraint.

5 Computational Results

To test the effectiveness of our ILP models we compared their results with the results obtained by a heuristic which was developed in the thesis of Schiefer [3] under the guidance of the first two authors.

The ILP models have been solved with ILOG CPLEX 9.0 on a 2 GHz PC. Real world test instances were provided by three different wards of the university hospital in Graz. To increase the variability, we created further test instances by varying demand and staff availability. To allow a comparison of schedules obtained by different algorithms, we constructed a valuation function that aggregates all aspects of a schedule into a single real value. The choice of the scalarization coefficients has been done based on extensive discussions with the personal of the involved wards.

The preliminary computational results for the multi-commodity model look promising. By using a time limit of 15 hours, we managed to solve about one third of the test instances to optimality. For the remaining instances the average gap between the value of the best found solution and the optimal value of the LP-relaxation was around 1%. For almost all instances the computed solutions were of a better quality than those derived by the heuristic. Further experiments are under way.

References

1. E.K. Burke, P. De Causmaecker, G. Vanden Berghe, H. Van Landeghem, “The state of the art of nurse rostering”, *Journal of Scheduling* **7**, 441–499, (2004).
2. P. Cappanera, G. Gallo, “A multicommodity flow approach to the crew rostering problem”, *Operations Research* **52**, 583–596, (2004).
3. E. Schiefer, “Computerized shift scheduling for hospital nurses”, Diploma thesis, Graz University of Technology, (2005).

Process Optimization and Efficient Personnel Employment in Hospitals

Gert Zülch, Patricia Stock, and Jan Hrdina

ifab-Institute of Human and Industrial Engineering, University of Karlsruhe,
Germany
jan.hrdina@ifab.uni-karlsruhe.de

1 Problem Area Hospital

1.1 Economic Conditions in Hospitals

The hospital field in Germany, with its 1.1 million employees and 62 billion Euro annual turnover (figures from 2001), represents a large and socially important field of activities [3, pp. 13]. If one considers the patients as “action object”, the work system (cf. [4, p. 81]) “hospital” is characterized by a number of particularities: First, treatment decision have to be made based on incomplete information and the treatment sequences must be carried out individually, meaning that they can only be planned to a limited degree. Furthermore, the complex treatment sequences must fulfil the treatment order reliably and efficiently. On the other hand, treatment processes must also take occupational health and safety and hygiene requirements of the medical and nursing personnel into account [10, p. 224].

The financing of hospitals (in part through case flat rates) is no longer oriented toward the actual costs of the individual hospitals [7, p. 142]. The fixed pricing system of the “German Diagnosis Related Groups” was changed as of 01.01.2004 so that, depending upon the patient, only a pre-determined amount is paid (§ 17 KHG - Hospital Financing Law). As a result, the costs aspect is gaining continuously in importance and hospitals must operate more cost-effectively in order to avoid losses.

Optimizing the personnel costs is of great significance for the economy of a hospital since they make up about two thirds of a hospital’s total costs [11, p. 389]. Consequently, a costs reduction is best achieved through a more efficient employment of the medical and nursing staff.

In order to be able to work more efficiently, flexible working time models, specially tailored to the respective hospitals, must be developed in order to better align the personnel capacity stock to the capacity requirements dictated by the arrival of patients.

1.2 Working Times in Hospitals

In the so-called “Jäger Verdict” from September 9, 2003, the European Court of Law declared that “the guideline 93, 104, EG of the European Council is to be

interpreted so that the stand-by duty of a doctor served by personal presence in the hospital represents in its entirety working time, even if said doctor is allowed, in times when he is not employed, to rest at his workplace.” Average weekly working times of more than 48 hours are, therefore, no longer admissible. Studies from the associations concerned assume that this creates a workforce deficit of 20,300 doctors and 12,900 other positions (all full-time)[12, p. 9].

However, due to their financial situation, hospitals are not capable of creating this number of new positions. Thus, the need for the larger personnel workforce must be reduced. The key to this problem lies in the efficient employment of personnel, in which personnel is only on duty when there is work to be done. The need for continuous medical care and monitoring is a further constraint for hospitals, which must be taken into account in working time models [5, p. 958].

In order to find a solution for the abovementioned problem, the *ifab*-Institute of Human and Industrial Engineering has developed a simulation-based procedure for the quantitative and efficient assessment of working time models in hospitals.

2 Simulation-Based Working Time Configuration for Hospitals

The personnel-oriented simulation of working systems has proven to be highly effective (cf. [13, p. 371]). Comprehensive, prospective information about the possible effects of planned changes can be gained using a simulation model capable of taking temporal work operations into account [2, p. 53]. This is not possible with conventional evaluation methods (e.g. value of benefit analyses, sensitivity analyses, list of pros and cons). In the following, the elements required for the simulation-based configuration of working times in hospitals will be presented.

2.1 Configuration Elements of Working Time Models

According to Bellgardt [1] the configuration elements of working time models can be classified into content and formal elements. The content elements refer to the object of the configuration, meaning the type of scheduling leeway allowed in the model. The formal elements define the borders within which the enterprise and the workers can use their leeway. The following presents the elements of both types (cf. [2, pp. 55], supplemented):

- Content configuration elements
 - duration
 - weekly working time in hours
 - minimum possible working time in hours
 - maximum possible working time in hours
 - placement
 - number of working time corridors (tuples with starting and ending times, as well as availabilities and possible compensation bonuses)
 - cycle period in days (time period after which the placement of the working times is repeated)
- Formal configuration elements

- Reference time frame,
- Implementation time frame in days,
- Compensation time frame in days, as well as
- Planning time frame in days.

2.2 Configuration Elements for Modelling Hospitals

In order to reconstruct hospitals realistically in a simulation model, one requires the following configuration elements (according to [6, pp. 94] and [2, pp. 102]):

- Patient structure and arrival
 - Patient type (characterized by the work operations to be carried out and their flow)
 - Patient arrival (characterized by the distribution of the arrival patterns per patient type)
- Functions
 - Direct functions, meaning functions to be carried out specifically for a patient (e.g. diagnosis or measuring of blood pressure)
 - Indirect functions, meaning functions that are not carried out for the patients directly, rather usually of preparatory or wrap-up nature or carried out for several patients (e.g. preparation for the distribution of medication, shift hand-off)
- Resources
 - Workplaces (e.g. treatment rooms, operating theatres)
 - Personnel (the workers employed and their qualifications)
 - Equipment (e.g. medical devices)
 - Material (e.g. medicine, medical supplies)

2.3 Assessment of Working Time Models

In order to assess whether a working time model is suitable for a hospital or not, an appropriate assessment concept is required. Here, the assessment concept based on degrees of goal attainment is used. A degree of goal attainment can take on a value between 0% and 100%, where 100% represents the ideal value of an index.

The assessment concept developed in this case is based on the assessment concept of Bogus [2, pp. 130] for the working time configuration in the service sector. In addition, also some hospital-specific figures concentrating on the patient-related key figures as well as on the key figures regarding the quality of care were defined:

- Hospital-oriented goal achievement
 - duration of patient stay (treatment path)
 - utilization of resources (personnel, facilities)
 - service rate
- Financial goal achievement
 - effective staff costs
 - patient handling costs
- Personnel-related goal achievement
 - degree of physical stress
 - degree of time stress
- Patient-related goal achievement
 - waiting times
 - uncompleted tasks

2.4 Heuristic Determination of the Best Flexible Working Time System

Using the modelling concept described above, various possible working time models can be modelled and examined with respect to their suitability for specific hospitals. In a simulation study the user of the simulation tool (i.e. the planner of working times) normally defines the alternative working time models using his experience. But the configuration of flexible working time models is rather complex which results in numerous alternatives or complementary applicable working time models.

Therefore, a special heuristic was developed which supports the determination of the best flexible working time system for each personnel type (e.g. doctors and nurses): It first determines the capacity requirements (for every personnel type) dynamically through simulation, under the assumption of an infinite capacity stock. Afterwards, a $(\mu+\lambda)$ -ES (for Evolution Strategies see e.g. [9]) is used in order to find the best working time system: μ randomly determined starting solutions generate λ offsprings, altering the content and formal configuration elements. The concrete working times for the employees are determined with the heuristic of Bogus [2, pp. 79]. The parents as well as the offsprings are simulated. In order to evaluate their fitness, the simulation results are assessed according to the benefit target system of the planner using an additive preference function. The μ fittest of the parents and the offspring will build the next generation. The $(\mu+\lambda)$ -ES terminates after a fixed number of generations or if a predefined level of fitness is achieved.

3 The Simulation Procedure *OSim-GAM*

The modelling concept described here was integrated into the object-oriented simulation procedure *OSim* (Object simulator, [8]), which was developed at *ifab*, thus creating the simulation procedure *OSim-GAM* (Object simulator, [2]). Furthermore, the heuristic for the generation of suitable working time models will also be integrated.

In *OSim-GAM*, patients are divided into various patient groups (outpatients, inpatients etc.) whose treatment sequences are represented using activity networks (network diagrams of the treatment operations). The activity networks are activated through so-called triggers, which possess an association to the activity networks. This can happen, for example when a patient arrives. Figure 1 presents an example of such an activity network for the patient type “intensive care patient” as well as the associated trigger.

Using *OSim-GAM*, changes to the treatment sequences can also be modelled. However, discussions with experts revealed that this feature was not suitable for the hospital in question.

4 Summary and Outlook

Due to changing general conditions, hospitals are now faced with new financial risks, which can only be counteracted by customized working time models. In this

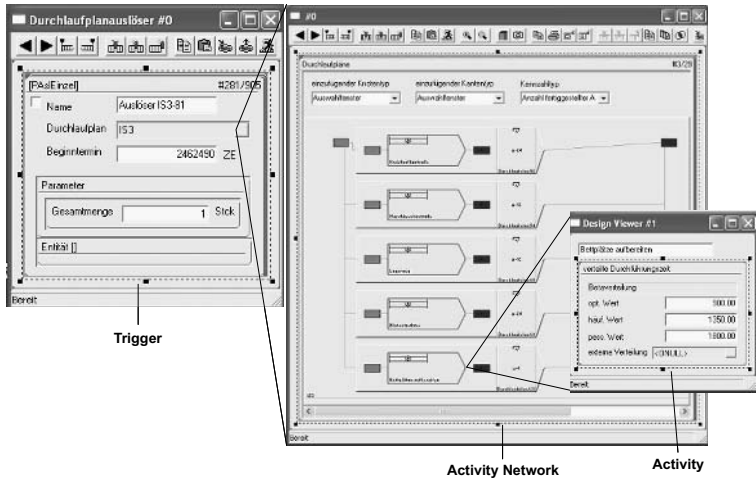


Fig. 1. Treatment path and trigger for the case of an intensive care patient

context, the simulation procedure *OSim-GAM* was extended through a hospital-specific modelling concept. Furthermore, a heuristic, with whose help appropriate working time models can be generated, was also integrated.

Simulation studies are currently being carried out in cooperating hospitals in order to validate the effectiveness of the procedure. For this, the effects of various working time models will be examined with regards to their goal attainment of the developed achievement concept.

Acknowledgments

The Authors would like to thank the German Research Association (Deutsche Forschungsgemeinschaft) for its support in the research work referred to in this paper. These endeavours are currently sponsored within a project dealing with the development of a simulation-based method of increasing the efficiency of personnel employment in hospitals (“Prozessoptimierung und effizienter Personaleinsatz im Krankenhausbereich - Gestaltung flexibler Arbeitszeitmodelle mit Hilfe der person-orientierten Simulation”).

References

1. Peter Bellgardt. *Ansatzmöglichkeiten des Arbeits- und Betriebszeitmanagements*, pages 85–104. Campus, Frankfurt/M., New York, 1990.
2. Thomas Bogus. *Simulationsbasierte Gestaltung von Arbeitszeitmodellen in Dienstleistungsbetrieben mit kundenfrequenzabhängigem Arbeitszeitbedarf*. Shaker, Aachen, 2002.
3. Statistisches Bundesamt Deutschland, editor. *Gesundheit: Ausgaben und Personal 2001*. Wiesbaden, 2003.

4. REFA-Verband für Arbeitsstudien und Betriebsorganisation, editor. *Methodenlehre der Planung und Steuerung - Teil 1*. Carl Hanser Verlag, München, 1991.
5. Ralf Gratias. Umsetzung des EuGH-urteils trifft auch die Pflegedienste. *Die Schwester - der Pfleger*, 41(11):956–961, 2002.
6. Max-Jürgen Heitz. *Ein engpassorientierter Ansatz zur simulationsunterstützten Planung von Personalstrukturen*. Shaker, Aachen, 1994.
7. Klaus Henning, Ingrid Isenhardt, and Clemens Flock. *Kooperation im Krankenhaus*. Huber, Bern, 1998.
8. Uwe Jonsson. *Ein integriertes Objektmodell zur durchlaufplanorientierten Simulation von Produktionssystemen*. Shaker, Aachen, 2000.
9. Zbigniew Michalewicz. *Genetic Algorithms + Data Structure = Evolution Programs*. Springer, 3rd edition.
10. o.V. Kurz berichtet. *Zeitschrift für Arbeitswissenschaft*, 57(5):224, 2003.
11. Herbert Schirmer. *Krankenhaus-Controlling*. expert Verlag, Renningen-Malmsheim, 1998.
12. Martin Waiger. Flexible tarifliche Regelungen notwendig. *ku-Sonderheft*, pages 8–9, 2004.
13. Gert Zülch, Jörg Fischer, and Uwe Jonsson. An integrated object model for activity network based simulation. In Jeffrey A. Joines, Russel R. Barton, Keebom Kang, and Paul A. Fishwick, editors, *Proceedings of the 2000 Winter Simulation Conference*, pages 371–380, 2000.

Logistics and Transport

Inventory Control in Logistic and Production Networks

Bernd Scholz-Reiter and Salima Delhoum

University of Bremen, Germany

`bsr@biba.uni-bremen.de`

`del@biba.uni-bremen.de`

Summary. The implementation of appropriate inventory control rules delivers the logistics operations of organizations a competitive advantage on the market place. This is also true for special types of networked manufacturing supply chains such as production networks which have complex topologies and increased complexity due to their logistic and production processes, information delays and the globalized markets. The paper builds the case of a production network and determines which inventory control policies applied in practice, order points and periodic-reviews, suit most the functions of the production network. Basically the production net consists of four manufacturing units that proceed to the joint development of products. In order to produce the four different items, the manufacturer orders from the other units as well as from the external supplier according to the exogenous customer demand. Although the case portrays a simplified instance of a production net, yet it already exhibits complex non linear dynamics inherent to such systems. The model is simulated with the system dynamics method which demonstrates the aptitude to describe distributed environments as long as product aggregation is permitted. The simulation results derived from the application of the inventory control rules on the production net will allow a detailed assessment of the suitability of each of the policies, in terms of the inventory costs criterion. A discussion on when to implement which inventory rule in respect to the functionalities of the production net will be pursued.

1 Inventory Control

Inventories generate costs whether they are physical inventories like warehouses or virtual ones such as trucks or trains that transport containers for example. It is important to keep these costs low otherwise production will be less profitable. Inventory control is about the minimization of the average cost per time unit while satisfying the incoming demand to the production system. This field started at the beginning of last century boosted by the growth of manufacturers' activities. Inventory control rules are described in the literature [1, 2, 3]. Modern inventory control saw the contributions of tools such as material requirements planning (MRP), its successor manufacturing resources planning (MRP II) and subsequently enterprise

resources planning (ERP), as well as Kanban cards to name a few. Recently RFID yields real time traceability and identification of products. Under this trend novel research examines autonomy in logistic processes [4].

In a production network manufacturers integrate and coordinate the general net production plan to their individual production planning to satisfy customer demand. When the internal complex dynamics of the network, such as production specifications and logistic processes, favour external disturbances, inventory oscillations will gain in intensity. A production network is a form of cooperation between manufacturers and the challenge consists in the integration of the various production plans. Former studies proposed the application of (a) load-oriented or order release control methods and (b) decentralized control loops [5]. These methods follow a black box solution approach.

In this paper the utilization of the inventory control rules is complemented by the modelling of the holistic network and its feedback structures through the continuous system dynamics methodology. Because it considers the entire network, the identification of leverage points where improvement is possible becomes an easier task. Continuous and periodic inventory control rules are applied to the system dynamics model of the logistic and production network in order to find a policy that reduces inventory costs. Costs are occasioned by products on hold and the backlogs of items ordered and not yet available. The latter is charged double the former as a penalty for the risk of customer loss. Under the continuous policies is meant an uninterrupted supervision of the inventory performed by the policies of (a) order point - order quantity (s, Q) and (b) order point - order up to level (s, S). On the contrary periodic review policies take into account constant order intervals. The ones investigated are (a) review period - order quantity (r, Q), (b) review period - order up to level (r, S), (c) review period - order point - order up to level (r, s, S), and (d) order inventory rule. In (d) the order O (1) is a function of the constant reacquisition time L (time between the order is placed and the supply is made), the safety factor z , the average demand d_{avg} (2) as well as its standard deviation during reacquisition d_{std} (3). The demand d_i is assumed to follow a normal distribution with σ_d the standard deviation. P is the length of a static period (one minute in this case).

$$O = (L \cdot d_{avg}) + (z \cdot d_{std} \cdot \sqrt{L}) \quad (1)$$

$$d_{avg} = \frac{d_i + \dots + d_{i+(L-1)}}{L}, \quad \text{where } i = 1, 2, 3 \dots \quad (2)$$

$$d_{std} = \sigma_d \cdot \sqrt{\frac{L}{P}} \quad (3)$$

2 Production Network

The case (1) describes a production network with four manufacturers, M_i , $i = 1 \dots 4$, who jointly produce items. The manufacturer has four parallel production lines L_{ij} , $j = 1 \dots 4$ and each of them produces one dedicated item P_{ij} . The product P_{ij} denotes product j manufactured by M_i . Each line is externally coupled with the production system of the other manufacturers. There is also internal coupling within a manufacturing unit (line L_{12} and L_{14}). A production line has a work-in-process

(WIP) that stores the products before production and a stock of manufactured lots. Every WIP has a minimum (null) value. The supply of raw materials for the WIP is assumed to be unlimited in capacity and it is done by an external supplier to the production line L_{ii} , $i = 1 \dots 4$. This line receives customer orders and passes them to lines L_{ij} . The line L_{ii} manufactures the finished items P_{ii} . These are then delivered outside the network to the customer. The line L_{ij} ($i \neq j$) receives the quantity ordered to produce P_{ij} from L_{ii} . The semi-finished parts P_{ij} , $i, j = 1 \dots 4$, ($j \neq i$), are jointly produced by two manufacturers: M_i (M_j) processes the part P_{ij} (P_{ji}) with M_j (M_i). The minimum lead time on L_{ij} is denoted T_{ij} . It is equal to the sum of the cycle time and the transport time from L_{ij} (M_i) to L_{ji} (M_j). T_{ij} is supposed to be constant.

Although the model is a simplified theoretical instance of a production network, similar re-entrant structures exist in real-life industries; one example is semiconductors. The choice is on four manufacturers because that renders the behaviour of the network more challenging, than with less manufacturers, in terms of the feedbacks and nonlinearities created by the production and logistic processes.

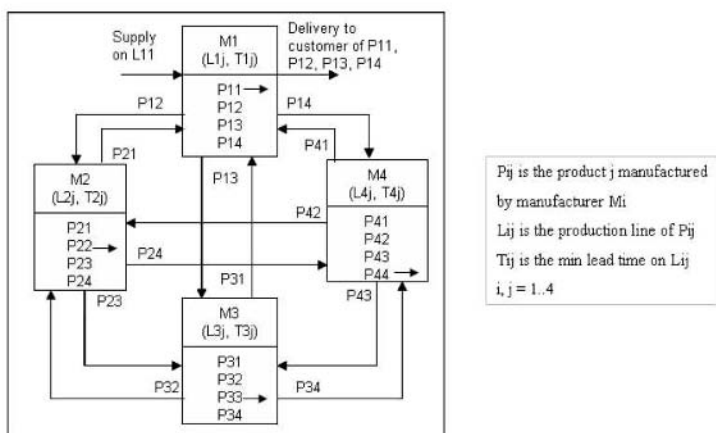


Fig. 1. Production network structure

The system dynamics model proposes a continuous modelling approach of the production system contingent on the continuous movement of flow and time. The former is argued by the fact that, despite the different products, the focus is on the aggregate lot. Indeed the interest is on the oscillatory behavior of inventories when disparate inventory control rules are applied. The cost function measures the intensity of these oscillations. Ordering costs are ignored. Instead a holding cost of 0.5 Euro per minute for a product in inventory is accounted for in addition to an out-of-stock penalty of 1.0 Euro per minute for an item in backlog. The simulation period lasts two thousand minutes and the simulation step is one minute.

3 Simulation Results

A step function was used to test the response of the model to a one hundred percent increase in customer demand for the products P_{13} and P_{31} at time one thousand minutes. This means that customer demand is quasi-constant. Even when this is the case, ample oscillations appear in supply chains because of material and information delays, feedback structures and nonlinearities [6]. The step function effect can be better seen on the graph of the network's average cost per minute. However, because on this graph the best inventory policy can be hardly distinguished, the graph of cumulative costs was rather preferred (2). After the iterative adjustment of the parameters of the inventory control rules (review period r , order point s , order up to level S , etc.) the models were simulated and the cumulative costs compared. The best inventory control policy for the production net is the order point - order up to level (s, S) rule. It has the capacity to adjust itself in case of uncertainty because both the order quantity and order interval change. At the opposite the review period - order quantity (r, Q) has the second highest costs. The reason is that the order interval and order quantity are both constant and hence, they render this rule inflexible. Consequently, the rule (r, Q) finds little practical applications. The review period - order point - order up to level (r, s, S) policy provides twenty five percent more costs than the best (s, S) rule at the end of the simulation period because of the constraint on the fixed order interval r . Indeed when r is null the periodic strategy (r, s, S) turns to the continuous inventory supervision (s, S). As expected the review period - order up to level (r, S) rule is more expansive than the order point - order quantity (s, Q) policy.

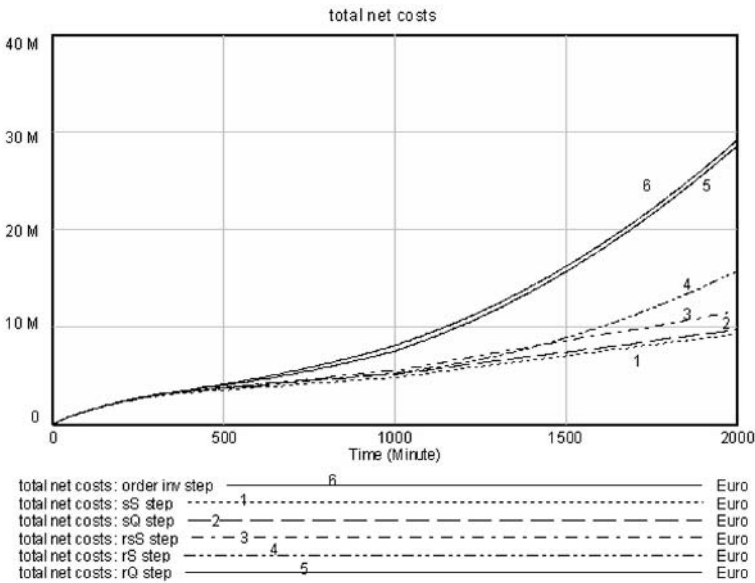


Fig. 2. Cumulative costs of the production net based on average costs per minute

4 Discussion and Conclusion

An inventory control rule that yields less inventory costs on the production network structure proposed was determined. This rule is the continuous order point - order up to level (s, S) strategy. A ranking of the inventory control rules from best to worst provides the following order: (s, S), (s, Q), (r, s, S), (r, S), (r, Q) and order inventory. Clearly, the choice of an appropriate rule depends not only on the costs, but also on how supply and demand are made: whether on a continuous or periodic manner. On this basis the simulation results obtained can help the scheduler choose one of the inventory control rules. It is important to generalize the present outcomes to other production network structures. In order to do that a qualitative approach is followed. It basically consists in the development of a learning medium from this logistic and production net. The idea behind it lies in the fact that decisions are mainly the task of humans; and therefore, in order to minimize the costs caused by risky decisions in such systems, schedulers need to develop some skills before actually placing orders. The Supply Net Game [7] emphasizes learning implicit skills of inventory control, collaboration and cooperation in distributed production environments. Decisions makers active in a wide range of industries can use the game. The industries targeted are those in which oscillations of orders and inventories alike occur such as the semiconductor and high-tech, commodities (beer, pampers...), automobile, aviation, chemicals or shipping and distribution. Experiments with the supply net game are still under way.

References

1. Gudehus T (1999) Logistik Grundlagen Strategien Anwendungen. Springer-Verlag, Berlin Heidelberg
2. Lödding H (2005) Verfahren der Fertigungssteuerung. Springer-Verlag, Berlin Heidelberg
3. Tempelmeier H (2005) Bestandsmanagement in Supply Chains. Books on Demand Verlag, Norderstedt
4. Scholz-Reiter B, Freitag M, de Beer C, Jagalski T (2005) Modelling dynamics of autonomous logistic processes: Discrete-event versus continuous approaches. *Annals of CIRP* 55: 413-416
5. Wiendahl H-P, Lutz S (2002) Production in networks. *Annals of CIRP on Manufacturing Technologies* 51(2): 573-586
6. Sterman J (2000) Business dynamics: Systems thinking and modelling for a complex world. Irwin McGraw-Hill
7. Scholz-Reiter B, Delhoum S (2006) The supply net game. Proceedings of the International System Dynamics Conference, July 23-27, Nijmegen, The Netherlands, <http://cgi.albany.edu/~sdsweb/sds2006.cgi?P179>

Vehicle and Crew Scheduling with Flexible Timetable

András Kéri and Knut Haase

TU-Dresden, Germany

`andras.keri@mailbox.tu-dresden.de`

`knut.haase@mailbox.tu-dresden.de`

1 Introduction

In this article we propose a new model for the simultaneous vehicle and crew scheduling problem occurring in the planning of urban mass transit systems. Our model incorporates the possibility of *trip shifting* in a given range in order to achieve a better solution.

Instead of defining the time windows for each trip individually, we define so-called *flexible groups* which are simply subsets of trips. We call a timetable which consists of flexible groups *flexible timetable*. Trips in the same flexible groups are shifted together.

The purpose of creating flexible timetable is twofold. In one hand, this could decrease the number of decision variables in the model. In the other hand, loss of service quality can be avoided: i) It keeps constant time headway between the trips in the same flexible group. ii) The waiting times of passengers does not change, if the trips of the lines whose connection are important are in the same flexible groups.

Each flexible group g has a time window $[d_{\min}^g, d_{\max}^g]$. The starting time of the trips in the same flexible group g should be changed with the same $t \in [d_{\min}^g, d_{\max}^g]$ value.

The structure of the paper is as follows: In section 2 the new model is presented. A heuristic solution method is described in section 3. Finally, in section 4 the computational result on smaller artificial instances is presented.

2 The Model

Our model is an extension to the model proposed by Haase *et al.* [3]. Let w be the set of trips, indexed by w . Each trip is divided into tasks (d-trip), which have to be performed by drivers. Let V represent the set of tasks, indexed by v . The set of flexible groups is notated by G , indexed by g . The set of trips is partitioned by flexible groups. W_g is the set of trips in the flexible group $g \in G$. Followed by the definition, $W = \bigcup_{g \in G} W_g$ and $W_{g_1} \cap W_{g_2} = \emptyset, g_1 \neq g_2, g_1 \in G, g_2 \in G$. Similarly, $V = \bigcup_{g \in G} V_g$ and $V_{g_1} \cap V_{g_2} = \emptyset, g_1 \neq g_2, g_1 \in G, g_2 \in G$. Let H represent the set

of the depot leaving time, indexed by h . Let U be the set of duty-types, indexed by u . The set of feasible duties are $\Omega^u, u \in U$, indexed by ρ .

The following binary parameters used in the model: $e_{v,t}^\rho$ is 1, if task v shifted by t time units is covered by duty ρ , otherwise 0. $f_{w,t}^\rho$ is 1, if duty ρ contains a driving movement ending at the start station of trip w shifted by t time units, otherwise 0. $g_{w,t}^\rho$ is 1, if duty ρ contains a driving movement starting at the end station of trip w shifted by t time units, otherwise 0. Let q_h^ρ equal to 1, if duty ρ contains a driving movement starting before time point $h \in H$, and ending after h , otherwise 0.

The binary variable $\theta_\rho^u, \rho \in \Omega^u, u \in U$ represents whether duty ρ is performed. The integer variable B is the minimal number of buses required to cover the schedule. The binary variable $X_{g,t}$ is a time-indexed variable, it is equal to one if the flexible group g shifted by t time units ($t \in [d_{\min}^g, d_{\max}^g]$).

The model of vehicle and crew scheduling with flexible timetable is the following:

$$\min c_B B + \sum_{u \in U} \sum_{\rho \in \Omega^u} c_\rho \theta_\rho^u \quad (1)$$

$$\sum_{u \in U} \sum_{\rho \in \Omega^u} e_{v,t}^\rho \theta_\rho^u = X_{g,t}, \quad \forall g \in G, \forall v \in V_g, \forall t \in [d_{\min}^g, d_{\max}^g] \quad (2)$$

$$\sum_{u \in U} \sum_{\rho \in \Omega^u} f_{w,t}^\rho \theta_\rho^u = X_{g,t}, \quad \forall g \in G, \forall w \in W_g, \forall t \in [d_{\min}^g, d_{\max}^g] \quad (3)$$

$$\sum_{u \in U} \sum_{\rho \in \Omega^u} g_{w,t}^\rho \theta_\rho^u = X_{g,t}, \quad \forall g \in G, \forall w \in W_g, \forall t \in [d_{\min}^g, d_{\max}^g] \quad (4)$$

$$\sum_{u \in U} \sum_{\rho \in \Omega^u} q_h^\rho \theta_\rho^u \leq B, \quad \forall h \in H \quad (5)$$

$$\sum_{t \in [d_{\min}^g, d_{\max}^g]} X_{g,t} = 1, \quad \forall g \in G \quad (6)$$

$$X_{g,t}, \theta_\rho^u \text{ are binary, } B \text{ is integer.} \quad (7)$$

The objective function (1) minimizes first the number of buses (here c_B is a large enough number), then the crew cost. (2) represent the task-covering equations. Task v shifted by t time units has to be covered if and only if the corresponding flexible group g is shifted with the same t value ($X_{g,t} = 1$). (3) and (4) are the so-called bus flow conservation equations. There should be exactly one driving movement ending at the start station of trip w shifted by t time units, and one starting at the end station of trip w shifted by t time units, if and only if the corresponding flexible group g is shifted with the same t value. (5) are the bus counting inequalities. (6) force that each flexible group is shifted by exactly one time value.

3 A Solution Approach

The number of feasible duties $|\Omega|$ is very large, direct enumeration of them yields an untreatable large linear problem. A column generation approach is more appropriate. Even with column generation, finding the optimal solution requires unacceptable running time. Because of this, we made the effort to find a heuristic method for solving this problem. Nevertheless, this method is also based on column generation. In the following subsections the main ideas of the method are described.

3.1 Solving the Main Problem

The linear relaxation of (1)-(7) represents the main problem of the column generation process. In each iteration of the process this relaxed problem is solved to optimality with the actual columns. The number of columns is controlled by two parameters, the minimal and maximal ratio of the number of columns to the number of rows. If the number of columns exceeds this value, columns with the highest reduced cost are deleted. To stabilize the dual variables, and thereby accelerate the convergence, an ϵ -bounded slack and surplus variables are used to perturb constraints (2)-(4), as Du Merle *et al.* proposed in [2].

After finding the optimal solution of the linear relaxation of (1)-(7), a round-up method is used to achieve an integer solution. Namely, the duty variable with the largest fractional value is fixed to 1.

3.2 Solving the Subproblem

For each duty-type $u \in U$ a similar driver network is defined as described by Haase *et al.* in [3]. There is only one difference. In [3] each task $v \in V$ represented by exactly one arc in each driver network. Here, each task $v \in V$ is represented by as many arcs as the number of possible shiftments of the corresponding flexible group. Hence, in our model a task covering arc represents a given task $v \in V_g$ shifted to a time point $t \in [d_{\min}^g, d_{\max}^g]$, $g \in G$. The labour regulations of each duty-type are represented by resource consumption through the arcs of the corresponding driver network.

Solving the subproblems is equivalent to find a resource constrained shortest path in each driver network. We used a dynamic programming algorithm proposed by Desrochers (1986) in [1]. However, in our approach this algorithm possibly yields an infeasible duty. This may occur when two tasks $v_1, v_2 \in V_g$ in the same flexible group $g \in G$ are covered by a duty, but the shiftment of the two tasks differs, namely $t_1 \neq t_2$. To avoid the generation such duties, a consistency check is performed in each time before a path is continued.

To speed up the algorithm, the number of labels in the nodes of the driver network are constrained by value c . In each node just the best c labels are processed. This greatly reduces the solving time of the subproblem. The experiments showed, that the solution is mostly equal to the optimal solution. The modified algorithm only not finds the optimal solution, when the pricing variables are near to zero.

4 Computational Results

We have implemented our algorithm in C++, compiled with the 64-bit version of GNU C++ compiler. The tests were run on an Sun Fire V40z computer with two 2.6GHz AMD Opteron processor and 16GB memory.

We have used two smaller test instances. Both of them had two lines, each line consisted of five station. We considered three stations on each lines as a relief point. The first instance had 32 trips, the second 133. We have used the same two labour rules as in [3]. We have assumed 20€ per hours for driving cost, and an additional 20€ fixed cost for each duty. We have defined seven flexible groups for each line in both instances. Details are given in figure 1.

Time period			Possible shiftments
0 : 00	-	7 : 00	-2, 0, 2
7 : 00	-	9 : 00	-2, 0, 2
9 : 00	-	16 : 00	-2, 0, 2
16 : 00	-	19 : 00	-2, 0, 2
19 : 00	-	24 : 00	-2, 0, 2

Fig. 1. Flexible groups

We have run the test first without using flexible timetable, then with using it. The results are shown in figure 2.

Inst. Algorithm	No. of buses	32-trip		No. of buses	133-trip	
		Crew cost	Time		Crew cost	Time
without TFF	8	901.3€	1min.	8	2695.5€	1 h.
with TFF	5	439.9€	32min.	5	1813.17€	23 h.

Fig. 2. Computational results

As we can see, our new approach yields a much better solution regarding the number of buses, and the crew cost. However it requires much more time to solve the problem.

References

- 1. Desrochers, M. (1986) La fabrication d’horaires de travail pour les conductors des autobus par une méthode de génération de colonnes, Ph. D. Dissertation, Université de Montréal, Montréal, In French
- 2. Du Merle, O., D. Villeneuve, J. Desrosiers and P. Hansen (1999) Stabilized Column Generation, *Discrete Mathematics* 194:229–237
- 3. Haase, K., G. Desaulniers, J. Desrosiers, (2001) Simultaneous Vehicle and Crew Scheduling in Urban Mass Transit Systems, *Transportation Science* 35(3):286–303

Lenk- und Ruhezeiten in der Tourenplanung

Asvin Goel and Volker Gruhn

Lehrstuhl für Angewandte Telematik/e-Business, Universität Leipzig, Germany
goel@telematique.eu
gruhn@ebus.informatik.uni-leipzig.de

Summary. Am 11. April 2007 werden in der Europäischen Union neue Vorschriften bezüglich der Lenk- und Ruhezeiten im Straßengütertransport in Kraft treten. Den neuen Vorschriften zufolge können Spediteure und Verloader für Verstöße der Fahrer haftbar gemacht werden. Da zudem Lenk- und Ruhezeiten einen signifikanten Einfluss auf Reisezeiten haben, ist eine Berücksichtigung der entsprechenden Regelungen bei der Tourenplanung unumgänglich. In diesem Beitrag wird gezeigt wie die neuen EU-Vorschriften in die Tourenplanung eingebracht werden können und das Vehicle Routing Problem mit Lenk- und Ruhezeiten (VRPLR) vorgestellt.

1 Einleitung

Am 11. April 2007 werden in der Europäischen Union neue Vorschriften bezüglich der Lenk- und Ruhezeiten im Straßengütertransport in Kraft treten [1]. Lenk- und Ruhezeiten haben i.d.R. einen signifikanten Einfluss auf Reisezeiten, die sich aus der reinen Fahrzeit und der Zeit für obligatorische Pausen ergibt. Den neuen Vorschriften zufolge, können zudem Spediteure und Verloader für Verstöße der Fahrer haftbar gemacht werden. Daher ist eine Berücksichtigung der entsprechenden Regelungen bei der Tourenplanung unumgänglich. Trotz ihrer Bedeutung in der Tourenplanung haben Regelungen bezüglich Lenk- und Ruhezeiten bisher nur wenig Berücksichtigung in der Literatur gefunden. [2] zeigen wie eine maximale Lenkzeit während einer Tour in Einfügeheuristiken berücksichtigt werden kann. [3] untersuchen ein Tourenplanungsproblem in welchem Mittagspausen und Nachtruhlen innerhalb vorgegebener Zeitfenster eingelegt werden müssen. Lenk- und Ruhezeitenregelungen des U.S. Department of Transportation wurden von [4] berücksichtigt, allerdings wird nicht betrachtet, dass Ruhezeiten eingelegt werden können bevor die maximale täglich Lenkzeit erschöpft ist. Solche „frühen“ Ruhezeiten, die notwendig sein können um Zeitfensterrestriktionen an folgenden Be- und Entladestellen einhalten zu können, wurden bisher nur in [5] berücksichtigt. In diesem Beitrag wird gezeigt wie Ruhezeiten und Lenkzeitunterbrechungen zwischen den Ruhezeiten in die Tourenplanung eingebracht werden können. In Abschnitt 2 werden wir zunächst die neuen EU-Vorschriften kurz beschreiben und dann zeigen, wie Lenk- und Ruhezeiten

berechnet werden können. In Abschnitt 3 wird dann das Vehicle Routing Problem mit Lenk- und Ruhezeiten vorgestellt.

2 Berechnung von Lenk- und Ruhezeiten

In diesem Beitrag werden folgende Vorschriften bezüglich der Lenk- und Ruhezeiten im Straßengütertransport betrachtet:

- Nach einer Lenkzeit (LZ) von viereinhalb Stunden ($t_{LZ} = 4.5$) muss ein Fahrer eine ununterbrochene Lenkzeitunterbrechung (LZU) von wenigstens 45 Minuten einlegen ($t_{LZU} = 0.75$), sofern er keine Ruhezeit einlegt.
- Nach einer täglichen Lenkzeit (TLZ) von maximal 9 Stunden ($t_{TLZ} = 9$) muss ein Fahrer eine ununterbrochene tägliche Ruhezeit (RZ) von mindestens 11 Stunden einlegen ($t_{RZ} = 11$).
- Während einer Lenkzeitunterbrechung oder täglichen Ruhezeit darf ein Fahrer frei über seine Zeit verfügen und es dürfen keine Arbeitszeiten (AZ) für Be- und Entladung oder anderen Tätigkeiten anfallen.
- Die wöchentliche Lenkzeit (WLZ) darf 56 Stunden nicht überschreiten ($t_{WLZ} = 56$).

Weitere Vorschriften, insbesondere Ausnahmeregelungen, werden in diesem Beitrag nicht berücksichtigt.

Wie Abbildung 1 zeigt, gibt es alternative Möglichkeiten Lenk- und Ruhezeiten einzuplanen, da Ruhezeiten eingelegt werden können bevor die maximale tägliche Lenkzeit erschöpft ist. Zur Berechnung möglicher Lenk- und Ruhezeiten sei das Label l_n für einen Knoten $n \in \mathcal{N}$ wie folgt definiert:

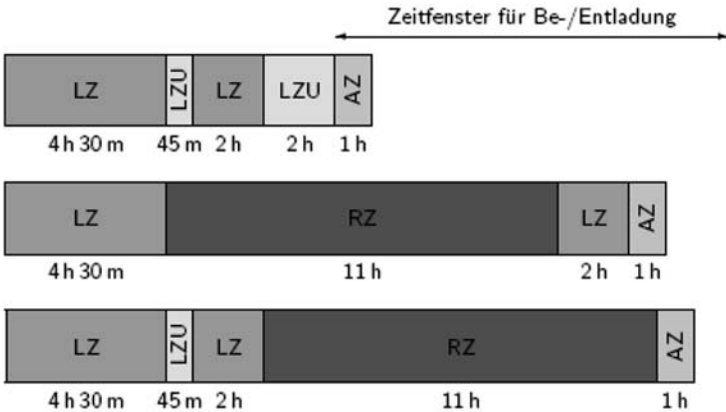


Fig. 1. Alternative Lenk- und Ruhezeiten

$$l_n = \begin{pmatrix} l_{n,1} \\ l_{n,2} \\ l_{n,3} \\ l_{n,4} \end{pmatrix} = \begin{pmatrix} \text{Ankunftszeit} \\ \text{wöchentliche Lenkzeit (WLZ)} \\ \text{tägliche Lenkzeit (TLZ)} \\ \text{Lenkzeit (LZ)} \end{pmatrix}$$

Ein Fahrzeug mit Label l_n erreicht den Knoten n zur Zeit $l_{n,1}$ und kann ihn frühestens zum Zeitpunkt $l_{n,1} + s_n$ verlassen, wobei s_n die benötigte Zeit für Be-/Entladung am Knoten n bezeichne. Der Fahrer darf noch maximal $t_{\text{WLZ}} - l_{n,2}$ bzw. $t_{\text{TLZ}} - l_{n,3}$ fahren bevor die nächste wöchentliche bzw. tägliche Ruhezeit eingelegt werden muss. Nach einer weiteren Lenkzeit von maximal $t_{\text{LZ}} - l_{n,4}$ muss die nächste Lenkzeitunterbrechung eingelegt werden.

Nicht alle der möglichen Label für einen Knoten müssen berücksichtigt werden und ein Label l_n *dominiert* ein anderes Label l'_n falls

$$l_{n,1} \leq l'_{n,1} \text{ und } l_{n,2} \leq l'_{n,2} \text{ und } l_{n,3} \leq l'_{n,3} \text{ und } l_{n,4} \leq l'_{n,4} \quad (\text{D1})$$

oder

$$l_{n,1} + t_{\text{LZU}} \leq l'_{n,1} \text{ und } l_{n,2} \leq l'_{n,2} \text{ und } l_{n,3} \leq l'_{n,3} \quad (\text{D2})$$

oder

$$l_{n,1} + t_{\text{RZ}} \leq l'_{n,1} \text{ und } l_{n,2} \leq l'_{n,2}. \quad (\text{D3})$$

Trivialerweise, muss ein Label, welches durch (D1) dominiert wird nicht berücksichtigt werden. Falls ein Label durch (D2) bzw. (D3) dominiert wird, so kann die Fahrt mit einer Lenkzeitunterbrechung bzw. Ruhezeit fortgesetzt werden, und danach würde (D1) gelten. Man beachte, dass keines der Label für die alternativen Lenk- und Ruhezeiten aus Abbildung 1 eines der anderen Label dominiert.

Wir zeigen nun wie mögliche Label l_m für einen Knoten $m \in \mathcal{N}$ berechnet werden können, wenn das betrachtete Fahrzeug von einem Knoten $n \in \mathcal{N}$ mit Label l_n zu dem Knoten m fährt. In diesem Beitrag nehmen wir an, dass ausschließlich Fahrten zwischen zwei wöchentlichen Ruhezeiten berücksichtigt werden müssen. Daher muss $l_{n,2} + \delta_{nm} \leq t_{\text{WLZ}}$ gelten, wobei δ_{nm} die reine Fahrzeit von n nach m bezeichne. Wir setzen zunächst $\mathcal{L} := \emptyset$ und

$$l_m := (l_{n,1} + s_n, l_{n,2} + \delta_{nm}, l_{n,3}, l_{n,4})^T.$$

Nun wird die rekursive Funktion, welche in Abbildung 2 beschrieben wird, durch `erweitere_Label( $l_m, \delta_{nm}$ )` aufgerufen. Zunächst wird die maximale ununterbrochene Lenkzeit Δ berechnet und das Label l_m sowie die verbleibende Fahrzeit δ aktualisiert. Falls $\delta = 0$ wurde der Knoten m erreicht und das Label l_m wird in die Menge $\mathcal{L}$ aufgenommen. Falls $\delta > 0$ und $l_{m,3} = t_{\text{TLZ}}$ wird eine tägliche Ruhezeit benötigt bevor der Fahrer weiterfahren darf und das Label l_m entsprechend angepasst. Falls die verbleibende Fahrzeit, die vor der nächsten Ruhezeit durchgeführt werden darf, nicht größer als die maximale Lenkzeit t_{LZ} ist, so wird ein neues Label l'_m mit Ankunftszeit $l_{m,1} + t_{\text{RZ}}$ erzeugt. Das Label l'_m ist unter Umständen dann wegen der geringeren täglichen Lenkzeit vorteilhaft, wenn aufgrund von Zeitfensterbeschränkungen an folgenden Knoten Ruhezeiten nicht beliebig eingelegt werden können. Dieses neue Label wird durch einen Aufruf `erweitere_Label( $l'_m, \delta$ )` erweitert. Schließlich wird die benötigte Zeit für eine Lenkzeitunterbrechung zu der Ankunftszeit des Labels l_m addiert und die Lenkzeit aktualisiert. Die Berechnung der Label wird dann mit der Berechnung der nächsten Lenkzeit weitergeführt.

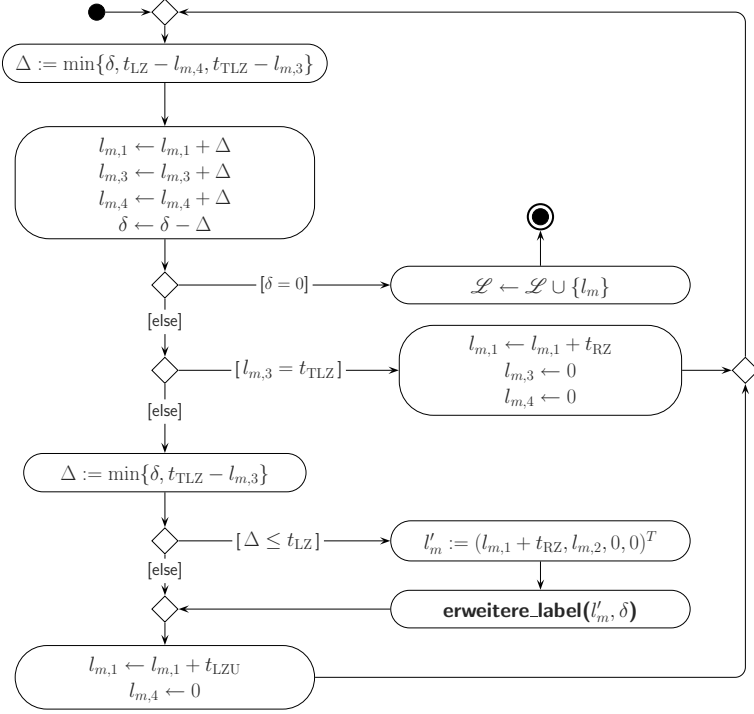


Fig. 2. Rekursive Funktion $\text{erweitere\_label}(l_m, \delta)$

Da die Ankunftszeit $l_{m,1}$ eines Labels $l_m \in \mathcal{L}$ vor dem Beginn des Zeitfensters $t_m^{\min}$ liegen kann, wird für alle $l_m \in \mathcal{L}$ eine Menge von möglichen Labels wie folgt bestimmt:

$$\mathcal{L}(l_m) := \left\{ \begin{aligned} &(\max\{t_m^{\min}, l_{m,1}\}, l_{m,2}, l_{m,3}, l_{m,4})^T, \\ &(\max\{t_m^{\min}, l_{m,1} + t_{LZU}\}, l_{m,2}, l_{m,3}, 0)^T, \\ &(\max\{t_m^{\min}, l_{m,1} + t_{RZ}\}, l_{m,2}, 0, 0)^T \end{aligned} \right\}$$

Nun sei $\mathcal{L}_m(l_n)$ eine Menge, die alle Label

$$l \in \{l'_m \in \bigcup_{l_m \in \mathcal{L}} \mathcal{L}(l_m) \mid l'_{m,1} \leq t_m^{\max}\}$$

enthält und in der kein Label durch ein anderes dominiert wird. Für ein Fahrzeug, welches von Knoten n mit Label l_n zu Knoten m fährt, sind somit alle Label $l_m \in \mathcal{L}_m(l_n)$ mögliche Label für Knoten m .

3 Das Vehicle Routing Problem mit Lenk- und Ruhezeiten

Das Vehicle Routing Problem mit Lenk- und Ruhezeiten ist das Problem Touren für eine Menge von Fahrzeugen $\mathcal{V}$ zu bestimmen, so dass jeder Kunde $n \in \mathcal{C}$ genau einmal innerhalb eines gegebenen Zeitfensters $[t_n^{\min}, t_n^{\max}]$ besucht wird, alle Touren den Lenk- und Ruhezeitenregelungen entsprechen und die Kosten für die Touren minimiert werden.

Für alle Fahrzeuge $v \in \mathcal{V}$ seien $n_{(v,1)}$ und $n_{(v,2)}$ Knoten die das Depot bezeichnen in dem alle Fahrzeuge stationiert sind. Für jedes Fahrzeug bezeichne $l_{n_{(v,1)}}$ das Label zu Beginn der Tour. Die Ankunftszeit am Ende der Tour muss für alle Fahrzeuge innerhalb des selben Zeitfensters $[t_{n_{(v,2)}}^{\min}, t_{n_{(v,2)}}^{\max}]$ liegen. Seien $\mathcal{D}_+ := \{n_{(v,1)} \mid v \in \mathcal{V}\}$, $\mathcal{D}_- := \{n_{(v,2)} \mid v \in \mathcal{V}\}$, $\mathcal{N} := \mathcal{C} \cup \mathcal{D}_+ \cup \mathcal{D}_-$ und $\mathcal{A} := \mathcal{D}_- \times \mathcal{D}_+ \cup (\mathcal{C} \cup \mathcal{D}_+) \times (\mathcal{C} \cup \mathcal{D}_-) \setminus \{(n,n) \mid n \in \mathcal{C}\}$. Für jede Kante $(n,m) \in \mathcal{A}$ bezeichne c_{nm} die Kosten der Kante ($c_{nm} = 0$ für alle $(n,m) \in \mathcal{D}_- \times \mathcal{D}_+$). Für jede Kante $(n,m) \in \mathcal{A}$ bezeichne x_{nm} eine binäre Variable welche angibt ob die Kante von einem Fahrzeug benutzt wird oder nicht. Für jeden Knoten $n \in \mathcal{C} \cup \mathcal{D}_-$ bezeichne l_n das zu bestimmende Label an dem Knoten. Das *Vehicle Routing Problem mit Lenk- und Ruhezeiten* (VRPLR) ohne Kapazitätsbeschränkungen ist dann

minimiere

$$\sum_{(n,m) \in \mathcal{A}} x_{nm} c_{nm} \quad (1)$$

unter den Nebenbedingungen

$$\sum_{(n,m) \in \mathcal{A}} x_{nm} = \sum_{(m,n) \in \mathcal{A}} x_{mn} \text{ für alle } n \in \mathcal{N} \quad (2)$$

$$\sum_{(n,m) \in \mathcal{A}} x_{nm} = 1 \text{ für alle } n \in \mathcal{N} \quad (3)$$

$$\text{für alle } (n,m) \in \mathcal{A} \text{ mit } n \in \mathcal{C} \cup \mathcal{D}_+ : x_{nm} = 1 \Rightarrow l_m \in \mathcal{L}_m(l_n) \quad (4)$$

$$x_{nm} \in \{0,1\} \text{ für alle } (n,m) \in \mathcal{A} \quad (5)$$

Die Zielfunktion (1) repräsentiert die akkumulierten Kosten über alle benutzten Kanten. Gleichung (2) sagt aus, dass die gleiche Anzahl Fahrzeuge einen Knoten erreichen und verlassen. Gleichung (3) sagt aus, dass jeder Knoten genau einmal erreicht wird. Durch (4) wird erreicht, dass alle Zeitfensterrestriktionen, sowie Lenk- und Ruhezeiten eingehalten werden. Letztlich wird durch (5) erreicht, dass die Entscheidungsvariable x_{nm} nur binäre Werte annimmt.

4 Schlussbemerkungen

In diesem Beitrag wurde gezeigt, wie die neuen Lenk- und Ruhezeitenregelungen, die am 11. April 2007 in der Europäischen Union in Kraft treten, in der Tourenplanung berücksichtigt werden können. Eine solche Berücksichtigung ist unumgänglich, da Lenk- und Ruhezeiten einen signifikanten Einfluss auf Reisezeiten haben und, den neuen Vorschriften zufolge, Spediteure und Verlader für Verstöße der Fahrer haftbar gemacht werden können. In Abschnitt 3 haben wir das Vehicle Routing Problem mit

Lenk- und Ruhezeiten vorgestellt. Heuristiken, die für das Vehicle Routing Problem with Time Windows entwickelt wurden, können oft so angepasst werden, dass sie auch für das VRPLR genutzt werden können. Hierfür müssen jedoch die Mengen der möglichen Label an jedem Knoten effizient verwaltet werden. Darüber hinaus kann die Zulässigkeit einer Änderung einer Tour nur dann ermittelt werden, wenn für alle Folgeknoten in der Tour die Labelmenge neu berechnet wird. Die höhere Planungsgüte sollte jedoch die zu erwartende längere Rechenzeit kompensieren.

References

1. Europäische Union. Verordnung (EC) Nr. 561/2006 des Europäischen Parlamentes und des Rates vom 15. März 2006 zur Harmonisierung bestimmter Sozialvorschriften im Straßenverkehr und zur Änderung der Verordnungen (EWG) Nr. 3821/85 und (EG) Nr. 2135/98 des Rates sowie zur Aufhebung der Verordnung (EWG) Nr. 3820/85 des Rates. Amtsblatt der Europäischen Union L 102, 11.04.2006, 2006.
2. Campbell A and Savelsbergh M. Efficient insertion heuristics for vehicle routing and scheduling problems. *Transportation Science*, 38(3):369-378, 2004.
3. Savelsbergh MWP and Sol M. DRIVE: dynamic routing of independent vehicles. *Operations Research*, 46:474-490, 1998.
4. Xu H, Chen Z-L, Rajagopal S, and Arunapuram S (2003). Solving a practical pickup and delivery problem. *Transportation Science*, 37(3):347-364.
5. Goel A and Gruhn V. Solving a Dynamic Real-life Vehicle Routing Problem. In: Haasis H-D, Kopfer H, and Schönberger J (eds), *Operations Research Proceedings 2005*, pages 367-372. Springer, 2006.

Transport Channel Selection

Jürgen Branke¹, Denis Häußler^{1,2}, and Christian Schmidt²

¹ Institute AIFB, University of Karlsruhe, Germany

`branke@aifb.uni-karlsruhe.de`

² Locom Software GmbH, Karlsruhe, Germany

`christian.schmidt@locom.de`

1 Introduction

The share of logistics and transportation in the overall cost of a product is increasing. Subsequently, there is a growing pressure in the logistics industry for optimization and cost reduction. In this paper, we consider the issue of transport channel selection, a problem that has been overlooked for a long time. Typical choices for transport channels include round-trips from and to the depot, one-way tours (often called “milkruns”), delivery over a hub, or fixed-cost delivery services (e.g. a postal service). The overall optimization problem consists of two interdependent sub-problems: assigning shipments to transport channels, and optimizing the routes within each channel.

In this paper, we propose an evolutionary algorithm (EA) with integrated savings heuristic to solve this two-level problem. Thereby, the EA is responsible for assigning shipments to transport channels. In order to evaluate an assignment within the EA, a vehicle routing problem is solved for each transport channel by means of a savings-based heuristic and subsequent 2-opt improvement. Empirical evaluation on five practical problem instances have shown that optimizing the transport channels by our algorithm can reduce the overall transportation cost significantly compared to the simple heuristics used in industry today.

The paper is structured as follows. In the next section, we specify the considered problem in more detail and give a short overview on how similar problems are typically solved in industry. Our proposed algorithm is described in Section 3. Empirical comparisons are given in Section 4. The paper concludes with a summary and outlook.

2 Transport Channel Selection

2.1 Problem Description

A typical transportation problem with channel selection can be described as follows: n customers have to be serviced from a unique depot. Each customer i has a given demand w_i for goods. There are different possibilities to service a customer, including:

1. Serving the customer by a milkrun.

A *milkrun* is defined as a tour starting at the depot, serving a number of customers, but *not* returning to the depot. Such tours are typically executed by a hired logistics partner. The way back is not payed, as the logistics provider can service other customers once the milkrun has finished.

2. Sending a dedicated truck directly to the customer.

This can be seen as a special form of milkrun.

3. Delivering over a hub.

If the demand of the customer is small but there are several customers in one region, the cheapest way of delivery may be to use a large truck to drive to a hub, and service the customers from the hub with a small truck.

4. Using a delivery service.

Usually, such delivery services charge a fixed cost depending on size, weight, and volume of the goods that can be looked up in a table.

The cost for servicing a customer through a delivery service or a dedicated tour are independent of how other customers are served. For milkruns and delivery over a hub, however, tours servicing several customers have to be formed, and thus the delivery cost depends on what other customers use the same transport channel. We assume that the cost for milkruns and tours from the depot is proportional to the driving distance. For the hub delivery, we additionally consider a fixed cost for using a hub.

A vehicle i used for delivery has a maximal capacity C_i and can drive a maximal distance D_i before it has to be back at the depot. The goal is to minimize the total transportation cost which is often equated with the total distance driven. A solution to the VRP is a collection of tours starting and ending at the depot, where each customer is assigned to exactly one tour, each tour has length at most D , and the total demand for each tour is at most C .

2.2 Related Work

To our knowledge, this problem has been largely neglected in the scientific literature so far. In industry, transport selection problems are usually solved manually with intuition and experience, or by rather simple heuristics. Variants of the transport channel selection problem are discussed in [4, 3], but without proposing a concrete algorithm.

Because we could not find any algorithm in the literature suitable for the transport channel selection problem as described above, we decided to base our empirical investigation below on a simplified version of the problem that has only two transport channels: milkruns and delivery services (and no hub delivery).

For this simplified problem, the following simple decision heuristics are commonly used in industry, and will be used for comparison:

Weight-based allocation (WA)

With this method, the user sets two weight limits w_{min} and w_{max} . All customers whose demand has a weight less than w_{min} are served via delivery service, all customers whose demand has a weight greater than w_{max} are serviced via a dedicated tour. For all other customers, a vehicle routing problem is solved. In the experiments below, we set $w_{min} = 2t$ and $w_{max} = 0.9C_{max}$ where C_{max} is the capacity of the vehicle.

First routing second improving (FRSI)

This method first solves a VRP with all customers. Then, in a second step it is attempted to improve the solution by moving some customers to the delivery service category. We consider two local improvement methods. In FRSI-t, all tours are examined in turn, and assigned to a delivery service as a whole if this is cheaper. In FRSI-c, we check for individual customers whether the savings by removing them from the tour is larger than the cost for the delivery service. If so, we move the customer to the delivery service, otherwise, we keep it in the tour. Because the removal of a customer changes the potential savings from removing another customer, we iterate through the set of customers until no more customer can be found whose assignment to the delivery service is beneficial.

3 Evolution-Based Transport Channel Selection

We propose to address the (general) transport channel selection problem by splitting it into two parts: first assigning transport channels to customers, and second solving the resulting VRPs at the main depot and the hubs.

For the assignment decisions, we propose to use an evolutionary algorithm (EA), a metaheuristic inspired by natural evolution. For an introduction to EAs, see e.g. [2]. To evaluate an assignment, we have to solve the corresponding VRPs, which is done by an embedded savings heuristic [1] with subsequent 2-opt.

The EA uses an integer encoding, with one variable for each customer specifying which of the 4 transport channels is to be used. If the decision is transportation over a hub, and if there are several hubs available, the customer is assigned to the closest hub. Crossover operator is simple uniform crossover, 90% of all offspring undergo mutation, which considers each variable in turn and with probability 25% assigns a new random value to the variable. The other 10% of the offspring employ a hill-climber to improve the assignment. The hill-climber first attempts to save cost by removing individual customers from existing tours and assigning them to the delivery service, then by considering for each customer assigned to the delivery service whether it can be beneficially inserted into an existing tour. Thereby, any improvements found are written back to the integer string representing the assignment (genotype).

The other parameter settings are more or less standard. We use a population size of 50, tournament selection, and generational reproduction with elitism (i.e. in every iteration, 50 new solutions are generated and replace the 50 old solutions, except that the best solution always survives to the next iteration).

For evaluating an individual/assignment, an embedded parallel savings heuristic is used to solve the corresponding VRPs at the main depot and each hub. For the milkruns starting at the depot, the savings heuristic was modified in a straightforward way to generate open tours not returning to the depot. After the savings heuristic, a local 2-opt hill-climber is used to improve the tour.

The algorithm terminates if it did not find an improvement for 50 iterations, or when it reached the maximum number of iterations allowed, 1000.

4 Empirical Evaluation

The 5 test problems used for comparison are based on real-world data. The number of customers and the total freight weight for each problem are given in Table 1, an example distribution of the locations for depot and customers is depicted in Figure 1.

For milkruns, we assumed a cost of 1.1 € per kilometer and a fixed cost of 50 € per stop. For using a delivery service, we used the tabulated rates from the German long-haul traffic-tariff (Güterfernverkehrstarif) minus 30%. For all vehicles, the maximal capacity was set to 15 t, and no distance restriction was used.

Table 1. Test problem characteristics

	Problem instance				
	1	2	3	4	5
No. of customers	311	291	252	253	186
Total weight [t]	1117,7	605,2	563,9	771,5	497,1

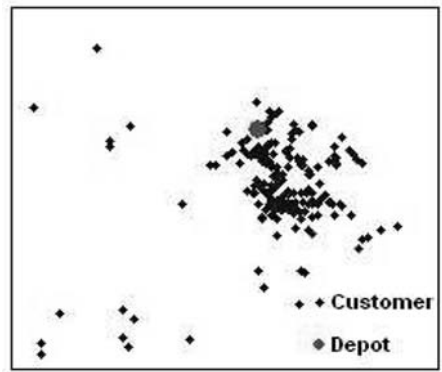


Fig. 1. Spatial distribution of customers for Problem 5

The results are summarized in Table 2. As expected, the simple weight-based allocation (WA) performs worst. Surprisingly, the FRSI method which tries to change complete tours (FRSI-t) to the delivery service is only slightly better (0.5%). If the local improvement is done on the level of individual customers (FRSI-c), savings increase to about 6%. Our EA-based heuristic clearly performs best, and is able to reduce transportation cost by on average 8.6% compared to the simple weight-based allocation.

Table 2. Results of the tested methods on the 5 problem instances. Results for the EA are averages over 100 runs $\pm$ standard error.

Problem	Method			
	WA	FRSI-c	FRSI-t	EA $\pm$ std. err.
1	46.4	44.0	47.3	43.5 $\pm$ 0.01
2	40.0	37.4	35.9	35.9 $\pm$ 0.01
3	33.4	30.7	33.4	29.5 $\pm$ 0.05
4	39.3	37.5	40.5	36.1 $\pm$ 0.01
5	33.2	31.8	34.5	30.9 $\pm$ 0.01
Avg.	38.5	36.2	38.3	35.2

5 Conclusion

This paper introduced the transport channel selection problem, an important real-world transportation problem which has been largely neglected in the scientific literature.

To solve this problem, we proposed an evolutionary algorithm (EA) to assign customers to transport channels. An integrated savings heuristic was used to determine the transportation cost for those transport channels that require routing.

Since we are not aware of other methods for the general transport channel selection problem, we tested our EA on a set of simplified real-world problems involving only the two transport channels “milkrun” and “delivery service”. On these problems, our algorithm compared favorably with the simple heuristics commonly used in industry. Thus, our algorithm not only performed better than the methods currently used in practice, but it also allows to handle more complex problems with more than two transport channels.

Since this work was only a first step in addressing the transport channel selection problem, there are ample avenues for future work. We are currently investigating the influence of the savings heuristic, and are experimenting with an improved savings method including a look-ahead. Also, we are planning to test our approach on related problem types such as the mixed fleet vehicle routing problem, which is investigated more thoroughly in the scientific literature. Other algorithmic approaches, be it metaheuristics like tabu search or fast local improvement heuristics, need to be developed.

References

1. G. Clark and J. W. Wright, *Scheduling of vehicles from a central depot to a number of delivery points*, Operation Research **12** (1964), 568–581.
2. A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing*, Springer, 2003.
3. C. Mues and S. Pickl, *Pickup and delivery Probleme mit Umlademöglichkeit - Ein Tourenplanungsproblem aus der Automobilzulieferindustrie*, Operations Research Proceedings 2001, Springer-Verlag Berlin, Heidelberg, 2001, pp. 33–40.

4. H. Wildemann and A. Niemeyer, *Logistikkostensenkung durch auslastungsorientierte Konsolidierungsplanung*, PPS-Management **4** (2002), 25–27.

A Sampling Procedure for Real-Life Rich Vehicle Routing Problems

Julia Rieck and Jürgen Zimmermann

Department for Operations Research, Clausthal University of Technology,
Germany
`julia.rieck@tu-clausthal.de`

Summary. In this paper we address a rich variant of the vehicle routing problem which occurs in real-life applications. Among other aspects we take into consideration time windows, simultaneous delivery and pick-up at customer locations, multiple use of vehicles, and timely allocation of vehicles to loading bays at the depot. In order to solve practical instances of the resulting real-life rich vehicle routing problem, efficient methods are required. For this reason, we present a sampling procedure, which is a multi-start algorithm that executes in each call a substantial extension of the well-known savings algorithm. Using a set of suitable benchmark instances, we assess the performance of the proposed sampling procedure.

1 Introduction

Most current problems faced by logistics operators have been studied for years, e.g. the vehicle routing problem (VRP), which is initially considered by Dantzig and Ramser [2]. The VRP consists of determining routes for vehicles of known capacity which are to operate from a single depot to supply a set of customers with known locations and known demands for a certain commodity. Routes for the vehicles are computed such that an objective function is minimized, e.g. the total distance traveled or the number of vehicles used. The problem has attracted a lot of attention in the academic literature, because it appears in a large number of practical situations and due to its complex structure, it is of theoretical interest. Many methods have been proposed for solving the VRP, either optimally or approximatively; we refer to [4] for a recent survey. In the past few years, the research community has turned to more complex, and rich variants of the VRP that occur in real-life applications. The family of those VRP variants is often identified as rich vehicle routing problems (RVRP). In the following we address a real-life RVRP, considering numerous constraints that are often found in reality (see e.g. [5]):

- heterogenous vehicles,
- a total working time for vehicles from departure to arrival at the depot,
- time windows for customers during which delivery and pick-up can occur,
- a time window for the depot,
- simultaneous delivery and pick-up at customer locations,

- multiple use of vehicles throughout the planning horizon, and
- timely allocation of vehicles to loading bays at the depot.

It should be noted that not all of these constraints may apply and therefore we present a sampling procedure which is able to solve this RVRP as well as generalizations.

2 Sampling Procedure

Due to the complexity of the problem at hand and the necessity to generate solutions in an appropriate time period, we use a *sampling procedure* to solve the problem approximatively. Our sampling procedure (cf. Section 2.1 – 2.4) is a multi-start algorithm that executes in each call a substantial extension of the savings algorithm proposed by Clarke and Wright [1].

The RVRP may be described as a graph theoretic problem. Let $G = (V, A)$ be a complete graph, where $V = \{0, \dots, n\}$ is the vertex set and A is the arc set. Vertices $i = 1, \dots, n$ correspond to the customers, whereas vertex 0 represents the depot. Two nonnegative weights d_{ij} and t_{ij} are associated with each arc $\langle i, j \rangle \in A$ which represent the distance and the travel time from vertex i to j , respectively. We assume that the distance and travel time matrices are symmetric and the triangle equation is satisfied; i.e. $d_{ik} \leq d_{ij} + d_{jk}$ for all $i, j, k \in V$. Each customer i is associated with a demand $d_i \geq 0$ that should be delivered and a demand $p_i \geq 0$ that should be picked-up. The service time $s_i > 0$ is the time which is necessary for the loading and unloading activities at customer i . A set of heterogenous vehicles is available at the depot and c_{max} is the capacity of the largest vehicle. To ensure that the feasible region is not empty, we assume that $d_i, p_i \leq c_{max}$ for $i = 1, \dots, n$. The vehicles have a total working time T_{max} which may not be exceeded by the duration dur_t of any tour t . The service at each customer i must start within an associated time window $[a_i, b_i]$, $a_i \leq b_i$. The time window $[a_0, b_0]$ represents the earliest possible departure from the depot and the latest possible arrival at the depot. We assume that each vehicle may perform at most two tours during the planning horizon.

2.1 Initialization

We start the sampling procedure with a simple solution that contains single customer tours $(0, i, 0)$ for each customer $i = 1, \dots, n$. For each tour $t = (0, i, 0)$ we determine a time window $[a(t), b(t)]$, where $a(t)$ is the earliest arrival time at the first customer of tour t and $b(t)$ is the latest arrival time at the first customer of t . Additionally, we define $w(t)$ as the waiting time and $c(t)$ as the core time of t , i.e. the sum of travel times and service times. For $t = (0, i, 0)$ we obtain $a(t) = a_i$, $b(t) = b_i$, $w(t) = 0$, and $c(t) = s_i$. Let ES_t , LS_t be the earliest and latest start time of tour $t = (0, i, 0)$ and l_t the time which is necessary for loading activities at the depot. Then we are able to calculate $ES_t = \max\{a_0, a(t) - t_{0i} - l_t\}$ and $LS_t = \min\{b(t) - t_{0i} - l_t + dur_t, b_0\} - dur_t$ for all tours t . In the initial solution each customer is served individually by a separate vehicle. Therefore, we assign the single customer tours to the existing vehicles, and if necessary we introduce fictitious vehicles. Later on (cf. Section 2.2), all non-required fictitious vehicles would be eliminated. When two routes $(0, \dots, i, 0)$ and $(0, j, \dots, 0)$ can feasibly be merged into a combined route $(0, \dots, i, j, \dots, 0)$, a distance saving $s_{ij} = d_{i0} + d_{j0} - d_{ij}$ for $i, j = 1, \dots, n$; $i \neq j$ is generated. We calculate all those savings and order them in a non increasing fashion.

2.2 Main Step

In the main step of our proposed algorithm we are trying to merge two tours t_1 and t_2 into a combined tour t' by connecting the first or the last customer i of t_1 with the first or the last customer j of t_2 . Moreover, it is possible to reverse t_1 or t_2 , if the resulting tour is feasible. The algorithm chooses, out of the set of potential combined tours, a feasible tour t' with maximum distance saving. Tour t' is feasible, if

- the capacity requirements $L(t') = \max_{i \in t'} L_i$ for tour t' are less or equal than c_{max} , where L_i represents the load after the visit of customer i ; i.e. $L_0 = \sum_{i \in t'} d_i$, $L_1 = L_0 - d_1 + p_1$, etc.,
- the duration $dur_{t'}$ of t' is less or equal than T_{max} ,
- customer j of t_2 can be reached timely, i.e. $a(t_1) \leq b(t_2) - t_{ij} - w(t_1) - c(t_1)$,
- t' can be executed during the depot opening time $ES_{t'} \leq LS_{t'}$, and
- an assignment algorithm that determines an assignment of the current set of routes to vehicles and loading bays, terminates successfully.

If all distance savings have been regarded, we will obtain a solution for the RVRP which is feasible, if it does not contain any fictitious vehicles.

The *assignment algorithm* tries to find an assignment of the current set of routes to vehicles and loading bays. Thereby, we assume that every vehicle is able to perform only two tours during the planning horizon. Firstly, we sort the tours in a non increasing order of capacity requirements and assign them in this order to the real vehicles and after that to the fictitious vehicles. If there is no tour just assigned to the current vehicle v , the vehicle will be able to execute the tour t' . If there is a tour s assigned to vehicle v , we have to check, if tour s allows the execution of tour t' . To this end we examine, whether tour t' can be assigned to the vehicle before tour s , perhaps with the utilization of the total float time $TF_s = LS_s - ES_s$ of s , or after tour s . If vehicle v is able to perform t' , then we have to check if there exists a loading bay for loading activities $l_{t'}$ and a loading bay for unloading activities $\bar{l}_{t'}$. We always stipulate that each tour t' starts at $ES_{t'}$ to guarantee an arrival at the depot on time. In case the current tour cannot be assigned to any loading bay, we allow waiting times at the depot. Only if there is a feasible assignment for all tours to vehicles and loading bays (cf. Fig. 1), the assignment algorithm will terminate successfully.

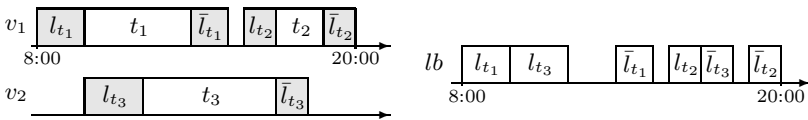


Fig. 1. Feasible assignment of t_1, t_2 and t_3 to vehicles v_1, v_2 and loading bay lb

Rational behind this algorithm is that, considering an appropriate sequence in which we merge tours, an optimal solution can be obtained (cf. Section 2.3).

2.3 Theoretical Background

One of the aims of this paper is to show the efficiency and the effectiveness of our sampling procedure. The algorithm is constructed in such a way that an existent

optimal solution can be reached, if we merge the tours in an appropriate sequence. We show this statement, starting with the capacitated VRP (CVRP) and then we add more and more constraints to the problem at hand.

The CVRP is an extension of the VRP in which a set of vehicles with different capacities is available at the depot. An optimal solution consists of a set of routes $T = \{t_1, \dots, t_r\}$ that cannot be merged without violating some constraints (triangle equation is satisfied). For each route $t_p \in T$ there exists a sequence in which we merge tours, beginning with the single customer tours, so that t_p is obtained. Due to the sorting of tours and vehicles in the assignment algorithm (cf. Section 2.2) we are able to find an assignment of $t_1, \dots, t_r$ to the real vehicles.

The VRP with time windows (VRPTW) is characterized by time intervals $[a_i, b_i]$ which are associated with each customer i . Due to the fact that each optimal solution satisfies these time windows, we only have to verify our calculation of $[a(t), b(t)]$, $w(t)$ and $c(t)$ for all tours $t \in T$. The core time, which is the sum of travel and service times, can be regarded as a constant value, because it is independent of the arrival times at the customers. Therefore, we are able to choose a fix reference point for each tour, e.g. the arrival at the first customer of a tour. Fig. 2 shows tours t_1, t_2 with reference points r_1 and r_2 before connecting customers i and j .

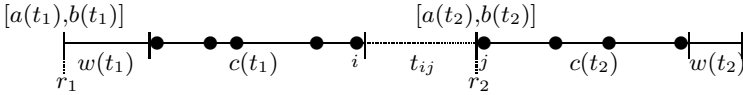


Fig. 2. Tours t_1 and t_2 with time windows, waiting times and core times

The waiting time of a combined tour t' is determined as follows

$$w(t') = \max\{a(t_2) - t_{ij} - c(t_1) - b(t_1), w(t_1)\} + w(t_2). \quad (1)$$

The latest arrival time at the first customer of t' is analogously defined as

$$b(t') = \min\{b(t_1), b(t_2) - t_{ij} - c(t_1) - w(t_1)\}. \quad (2)$$

For the earliest arrival time at the first customer of tour t' , it holds that

$$a(t') = \max\{a(t_1), a(t_2) - t_{ij} - c(t_1)\}, \text{ if } w(t') = 0, \text{ and } b(t') \text{ otherwise.} \quad (3)$$

Equation (3) ensures that $a(t') \leq b(t')$. This becomes clear by considering the tree different cases in which we specify $a(t') := b(t')$.

- If $w(t_1) = w(t_2) = 0$ and $w(t') > 0$ then we conclude $a(t_2) - t_{ij} - c(t_1) > b(t_1)$ from (1). It follows with (3) that $a(t') > b(t_1) \geq b(t')$.
- If $w(t_1) > 0$ then $a(t_1) = b(t_1)$ holds. We estimate $a(t') = \max\{b(t_1), a(t_2) - t_{ij} - c(t_1)\} \geq b(t_1) \geq b(t')$ with the aid of (2).
- If $w(t_2) > 0$ then $a(t_2) = b(t_2)$ holds and $a(t') = \max\{a(t_1), b(t_2) - t_{ij} - c(t_1)\} \geq b(t') + w(t_1) \geq b(t')$; cf. (2).

The VRP with multiple use of vehicles (VRPMU) is a CVRP in which the vehicles are able to perform multiple tours during the planning horizon. For our purposes we assume that the vehicles can execute two tours. Due to the sorting of tours, we firstly try to assign the tour with the largest capacity requirements to a real vehicle, then the tour with the second largest capacity requirements, etc. If we sort the real vehicles in a non increasing order of capacities, this will be the best method in many cases. But, if the optimal solution consisted of four tours with the following data:

$$\begin{aligned} t_1: L(t_1) = 60, [a(t_1), b(t_1)] &= [8, 10] & t_2: L(t_2) = 59, [a(t_2), b(t_2)] &= [14, 16] \\ t_3: L(t_3) = 58, [a(t_3), b(t_3)] &= [8, 14] & t_4: L(t_4) = 55, [a(t_4), b(t_4)] &= [10, 15], \end{aligned}$$

we would never obtain the optimal solution in which v_1 ($cap_{v_1} = 60$) performs the tours t_1, t_4 and v_2 ($cap_{v_2} = 59$) the tours t_2, t_3 . Therefore, we sort the real vehicles in a non increasing fashion of capacities and randomize this sorting after some iterations.

Our last extension of the VRP is the timely allocation of vehicles to loading bays at the depot. After determining a vehicle for a tour t , we have to check if there exists a loading bay for loading activities and a loading bay for unloading activities. We always terminate the start of a tour t to ES_t , but to avoid an early rejection of a tour, we allow waiting times at the depot (cf. Fig. 3).

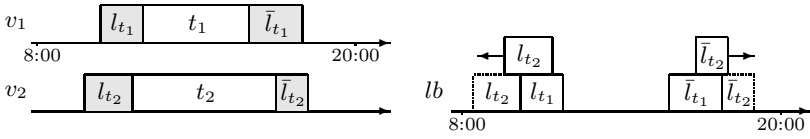


Fig. 3. Feasible assignment of t_1, t_2 and t_3 to vehicles v_1, v_2 and loading bay lb

In Section 2.4 we describe a multi-start method that considers different sequences in which we merge tours, so that an optimal solution is normally generated, if a high computational time is accepted.

2.4 Multi-Start Method

In the following we present a multi-start method which uses savings in a randomised sequence in the algorithm given in Section 2.1 – 2.2 to generate many different solutions. We vary the sequence in which we merge the tours with the “multi-pass regret-based biased sampling” proposed by Drexler [3]. In each call of the sampling procedure we determine for each saving $s \in S$ a basic parameter $\psi(s) = s - \min_{h \in S} h$ which is the absolute difference between the saving s and the smallest saving h . In a random experiment each saving is selected with the probability

$$\Psi(s) = \frac{(\psi(s) + 1)^\alpha}{\sum_{s' \in S} (\psi(s') + 1)^\alpha}$$

with $\alpha \in [0, \infty)$. For $\alpha \rightarrow \infty$, we choose the saving with the largest value and for $\alpha \rightarrow 0$ each saving has the same probability for choice. In every call of the sampling procedure we generate a solution, and we store the best one which could be found. The procedure terminates once a stop criteria is satisfied, e.g. some prescribed computational time or iteration number.

3 Preliminary Computational Results

In order to test the behavior of the proposed sampling procedure, we consider instances RZ for the RVRP which are randomly generated, taking into account control parameters to reproduce real-life situations. Additionally, we examine the Solomon instances (cf. [6]) for the VRPTW, because their design highlights several factors that affect the behavior of routing algorithms, e.g. geographical data, percentage of time-constrained customers, tightness and positioning of time windows, etc. Instances RZ are compared with optimal solutions computed with CPLEX and the

Solomon instances are compared with the optimal solutions from Solomon’s web page (cf. [7]). Our sampling procedure terminates after passing 100 iterations for each $1 \leq \alpha \leq 40$, respectively. The average distance deviations (in %) are shown in Tab. 1 and Tab. 2.

Table 1. Results for instances RZ

Instances	RZ
$n = 10$	0,00
$n = 15$	0,04

Table 2. Results for the Solomon instances

Instances	C1	C2	R1	R2	RC1	RC2
$n = 25$	0,36	0,39	0,62	0,70	3,74	2,55
$n = 50$	0,45	1,29	2,85	5,01	11,26	18,16

For instances RZ the average distance deviation is 0,02% and for the Solomon instances the average distance deviation is 3,95%.

4 Conclusion

We have developed a sampling procedure to solve a large range of real-life RVRPs. The algorithm is suitable for practical applications, because it is able to find a good, feasible solution in an appropriate time period. Normally, the algorithm can find an optimal solution, if a high computational time is allowed. Important areas of future research are, e.g. the implementation of our random number generator and the computation of the extended Solomon instances proposed by Gehring and Homberger.

References

1. Clarke G, Wright JW (1964) Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research* 12: 568–581
2. Dantzig GB, Ramser JH (1959) The truck dispatching problem. *Management Science* 6:80-91
3. Drexel A (1991) Scheduling of project networks by job assignment, *Management Science* 37:1590-1602
4. Toth P, Vigo D (2002) *The vehicle routing problem*. Siam, Philadelphia
5. Rieck J, Zimmermann J (2006) Tourenplanung mittelständischer Speditionsunternehmen in Stückgutkooperationen. In: Haasis H-D, Kopfer H, Schönberger J (eds) *Operations Research Proceedings 2005*. Springer, Berlin et al.
6. Solomon M (1987) Algorithm for the vehicle routing and scheduling problems with time window constraints, *Operations Research* 35:254-265
7. <http://web.cba.neu.edu/~msolomon/problems.htm>

Market-Oriented Airline Service Design

Cornelia Schön

Institute for Economic Theory and Operations Research, University Karlsruhe,
Germany

`schoen@wior.uni-karlsruhe.de`

Summary. The decision of an airline about its service offering is a challenging task as it involves various decisions at the interface of operations and marketing: which origin-destination (OD) markets to serve, over which routes, and at which departure times (schedule design), at which price and other ticket conditions (pricing/fare product design), and what aircraft type to assign to each of these flights (fleet assignment). These decisions are highly interdependent with regard to their profit impact: on the one hand, the fleet schedule fixes a large part of the cost; on the other hand, schedule, price and fare conditions are the most important factors influencing passenger's choice and thus revenue. While schedule- and fare-related decisions are often treated in isolation in airline planning, profit maximizing service design should encompass the simultaneous determination of all features in the service package that drive profit. We develop a market-oriented model for airline network service design integrating flight schedule design, fleet assignment and pricing. Under suitable assumptions, the model is a mixed-binary problem with concave objective function and linear constraints that can be solved exactly by standard techniques. The optimal solution obtained at the strategic level can be used as input for operational revenue management models providing an interface for hierarchical decision making.

1 Introduction

Profit-oriented airline service design involves many interdependent decisions at the interface of operations and marketing, in particular schedule planning and pricing, defining to a large extent the ultimate product of an airline from the customer's point of view. Due to complexity and problem size, airlines usually treat schedule- and marketing-related decisions in isolation and further decompose the whole schedule planning process into a set of four core problems that are solved sequentially: schedule design, fleet assignment, maintenance routing, and crew scheduling [2], [7], [10]. Even these smaller core problems present significant modeling and optimization challenges and for most of the airlines schedule design is still a manual process due to complexity and finally to unaccounted marketing requirements in existing models.

From a market-oriented perspective it would be desirable to simultaneously determine schedule- and fare-related criteria driving profit. This integration is not only motivated by the quest for global optimization but also because 1.) recent advances in forecasting demand and modeling choice behavior at the itinerary/fare class level allow such integration for real-world applications, 2.) contrary to expectations, integrating pricing into schedule design decisions does not necessarily come along with an increase in computational complexity, and finally 3.) a direct interface to revenue management models is established. Before we turn to the development of the integrated model, we first review the relevant literature in the area of schedule design/fleet assignment, pricing and demand modeling.

Schedule Design. While the literature on fleet assignment is vast and decision support tools for fleetings are common [8], there are only a few articles on schedule design optimization. Approaches to profit-oriented schedule design or iterative schedule design/fleet assignment are described in, for example, [5] and [12], and for a survey [6]. More recently and most relevant to our approach, [9] developed a mixed integer linear program for integrated schedule design and fleet assignment (ISD-FAM) with the objective to maximize schedule profitability based on an incremental approach. To account for recapture and network effects caused by schedule modifications, unconstrained itinerary-specific average demand for an initial schedule is adjusted by approximate demand correction terms. Demand is not modeled as a closed-form function depending on the controls but instead selected values are estimated with a schedule evaluation model and updated iteratively as needed when the schedule has significantly changed during the optimization. Prices are predetermined and enter the model as mean itinerary fare values averaged across booking classes. The assumption of fixed prices is common in schedule design models with the exception of [5]. They propose a nonlinear model and a heuristic for the simultaneous determination of profit-maximizing flight schedules and itinerary prices in a hub-and-spoke network. Demand is incorporated as a multinomial logit function influenced by price and schedule preference. There is only one size of aircraft and thus no fleet assignment.

Pricing. On the other hand, there are numerous approaches in the literature for pricing airline services given a fleet schedule [13]. For the purpose of integrating pricing with schedule design, we are interested in more strategic network pricing models and therefore focus on the deterministic multi-product multi-resource case. A suitable standard model is given in [13], ch. 5.4.1, that can be considered as a special case of our integrated approach when all fleet assignment decisions are fixed.

Demand. Pricing models assume that the decision maker has information about the price-demand relationship in terms of a suitable multi-product demand function (see [13], ch. 7.3.3). Today, both practitioners and researchers are aware of the need to account for customer behavior in traditionally operations-driven airline planning problems likewise. Discrete choice models, especially the Multinomial Logit (MNL) model, have been applied for a long

time for explanatory and predictive purposes to model (air) travel demand because they are a natural way to capture demand interdependencies [3]. Recent progress has been made in forecasting passenger behavior at the detailed booking class/itinerary level [1], [11]. In [4] for example an aggregate itinerary market share MNL model is estimated for all city pairs in the US with various independent variables: level-of service, connection quality, air carrier and carrier market presence, fares, aircraft size and type, and time of day. Despite the increasing availability of demand data and estimation techniques, such demand models have rarely been incorporated directly into normative schedule design models with [5] being an exception.

Outline and Main Results. Motivated by the discussion above we develop an approach for simultaneous schedule design, fleet assignment and pricing. The model presented in next section is a step towards global optimization in airline network service design where the most important factors driving profit are determined jointly. Under suitable assumptions, our formulation is a mixed-binary maximization problem with concave objective function and linear constraints that can be solved exactly by standard techniques. Thus, contrary to expectations, integrating pricing into schedule design and fleeting decisions does not necessarily come along with an increase in computational complexity. The optimal solution of the integrated model can be used directly as parameter input to revenue management models narrowing the gap between operations and marketing as will be further discussed in section 3, together with possible extensions.

2 Model

There are a number of terms and assumptions underlying our integrated model commonly used in the literature, namely: *Markets* are defined by an OD pair, assumed to be independent of one another. A *flight (leg)* is a scheduled and direct connection between a particular OD. An *itinerary* consists of one or more connecting flights that passengers can travel. A *fare product* is a combination of itinerary and booking class (ODF) provided that demand data is available at this detailed level; otherwise it is an itinerary aggregated over booking classes as in [9] or [5]. The decision about including a fare product in the schedule is not modeled by using binary variables but instead indirectly controlled by the pricing decision assuming that not offering an itinerary corresponds to raising the price to a sufficiently high level that turns off demand.

Schedule. Similar to [9], [5] and [4] we assume a daily schedule. Our approach to schedule design is incremental, i.e. the model takes as input a prespecified list of flight legs composed of mandatory and optional flights for deletion or addition [9]. The underlying network is a time line network.

Demand. Customers may be segmented by their preference for different service characteristics and simultaneously by the time of their reservation request which is typical e.g. for business/leisure passengers. Thus we can charge

different prices for different customers or, in other words, provide alternative prices as input to revenue management models for dynamically pricing the same product over time during the subsequent booking process. For each customer segment s , we assume to have a multi-dimensional demand function $d_{js}(\mathbf{p}_s)$ at fare product level j depending explicitly on prices $\mathbf{p}_s = (p_{sj})_{j=1..n}$ and implicitly on schedule-related and possibly other product characteristics (as well as on competitive products). Furthermore, we make some regularity assumptions shared by most common demand functions [13]: (a) the demand function has an inverse denoted by $p_{sj}(\mathbf{d}_s)$ – this allows to consider demand rather than price as the decision variable which avoid nonlinearities in the constraints; (b) $d_{sj}(\mathbf{p}_s)$ is non-negative, strictly decreasing in p_{sj} with $d_{sj}(\mathbf{p}_s) = 0$ for $p_{sj} \rightarrow \infty$, differentiable and bounded from above $\forall \mathbf{p}_s \geq \mathbf{0}$; (c) the revenue $d_{sj}p_{sj}(\mathbf{d}_s)$ is concave on the set of achievable demand vectors. In an accompanying paper we show how the standard and the nested MNL and other common demand models can be fitted into this framework. The following notation is used:

$\mathcal{S}$:	set of customer segments, e.g. business, leisure
$\mathcal{K}$:	set of fleet types, indexed by k
$\mathcal{A}$:	set of airports indexed by o
$\mathcal{J}$:	set of potential fare products at itinerary level, indexed by j
$\mathcal{L}$:	set of flight legs in the flight schedule, indexed by l
$\mathcal{T}$:	sorted set of all event (departure or availability) times at all airports, indexed by t_n . The event at time t_n occurs before the event at time t_{n+1} . Suppose $ \mathcal{T} = m$; then t_1 is the time associated with the first event after the count time and t_m the time associated with the last event before the next count time.
$\mathcal{N}$:	set of nodes in the time line network indexed by $\{k, o, t_n\}$
$\mathcal{L}^c(k)$:	set of flights passing the count time when flown by fleet type k
$\mathcal{I}(k, o, t)$:	set of inbound flight legs to node $\{k, o, t_n\}$
$\mathcal{O}(k, o, t)$:	set of outbound flight legs from node $\{k, o, t_n\}$
CAP_k :	number of seats available on aircraft of fleet type k
N_k :	number of aircraft in fleet type k
δ_{lj} :	indicator with $\delta_{lj} = 1$ if product j includes leg l , 0 otherwise
c_{sj} :	variable costs for a passenger of segment s with fare product j
$p_{sj}(\mathbf{d}_s)$:	inverse price-demand function for segment s and fare product j
d_{sj} :	demand of segment s for fare product j (var.)
f_{kl} :	1 if flight leg l is assigned to fleet type k , 0 otherwise (var.)
$y_{kot_n^+}(t_n^-)$:	number of fleet type k aircraft on the ground at airport o immediately after (before) time t_n . If t_1 and t_2 are the times associated with adjacent events, then $y_{kot_1^+} = y_{kot_2^-}$ (var.)

The integrated model for schedule design, fleet assignment and pricing can now be formulated as:

$$\begin{aligned} \max \quad & \sum_{s \in \mathcal{S}} \sum_{j \in \mathcal{J}} (p_{sj}(\mathbf{d}_s) - c_{sj}) d_{sj} - \sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}} f_{kl} C_{kl} \\ & \sum_{k \in \mathcal{K}} f_{kl} = 1 \quad \forall l \in \mathcal{L}^m, \quad \sum_{k \in \mathcal{K}} f_{kl} \leq 1 \quad \forall l \in \mathcal{L}^o \end{aligned} \quad (1)$$

$$y_{kot^-} + \sum_{l \in \mathcal{I}(k,o,t)} f_{kl} - y_{kot^+} - \sum_{l \in \mathcal{O}(k,o,t)} f_{kl} = 0 \quad \forall (k,o,t) \in \mathcal{N} \quad (2)$$

$$\sum_{o \in \mathcal{A}} y_{kot_m^+} + \sum_{l \in \mathcal{L}^c(k)} f_{kl} \leq N_k \quad \forall k \in \mathcal{K} \quad (3)$$

$$\sum_{s \in \mathcal{S}} \sum_{j \in \mathcal{J}} \delta_{lj} d_{sj} \leq \sum_{k \in \mathcal{K}} CAP_k f_{kl} \quad \forall l \in \mathcal{L} \quad (4)$$

$$f_{kl} \in \{0,1\} \quad \forall k \in \mathcal{K}, l \in \mathcal{L}, \quad d_{sj} \geq 0 \quad \forall s \in \mathcal{S}, j \in \mathcal{J} \quad (5)$$

$$y_{kot^-}, y_{kot^+} \geq 0 \quad \forall (k,o,t) \in \mathcal{N} \quad (6)$$

The objective is to determine the fare products with corresponding prices to be included in the schedule and an assignment of flights to fleet types such that the total contribution from the schedule over all customer segments is maximized. The constraints (1) ensure that each mandatory (optional) flight is covered by exactly (at most) one fleet type. Constraints (2) are the balance constraints for each node in the time line network, while (3) bounds the fleet-specific number of aircraft on the ground and in the air from above to the available capacity. The capacity constraints in (4) ensures that (a) total demand for a leg does not exceed the capacity assigned to that leg, and (b) that a fare product can not be offered (i.e. must carry zero demand) unless all flight legs in the underlying itinerary are fleeted. The decision variables are defined in (5) and (6).

3 Discussion and Conclusion

Under our reasonable regularity assumptions regarding demand the optimization model is a mixed-binary problem with concave objective function and linear constraints. In contrast to the case when prices are predetermined exogenously[9], we can avoid schedule-related binary variables and instead introduce continuous price (demand) variables facilitating the optimization significantly. The number of binary variables in our integrated approach is the same as in the basic fleet assignment model that is relatively easily solvable even for large flight networks by commercial integer programming solvers [8]. If the binary variables are relaxed, the resulting continuous concave maximization problem is quiet easy to solve exactly.

The optimal solution (itineraries, capacities, and a single or several alternative prices for each fare product) can be used directly in revenue management. Many revenue management models assume that – in addition to capacities and itineraries resulting from schedule planning – prices for fare classes are predetermined and provided by some exogenous marketing decision making process – either as a single price for each fare class in capacity allocation models or as a predefined set of alternative fares in dynamic pricing models with discrete prices. Thus, we can establish a direct interface between schedule design/fleet assignment, pricing and revenue management.

There are several straightforward extensions to our base model, e.g. considering physically different compartments, capacity constraints of competitors, cross-effects between markets and between days of the week.

References

1. S. Algers and M. Beser. Modeling choice of flight and booking class. *Intern. Journ. of Services, Techn. & Mgmt.*, 2(28):45, 2001.
2. C. Barnhart and A. Cohn. Airline schedule planning: Accomplishments and opportunities. *MSOM*, 6(1):3–22, 2004.
3. M. Ben-Akiva and S. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA, 1985.
4. G. Coldren, F. Koppelman, K. Kasturirangan, and A. Mukherjee. Modeling aggregate air-travel itinerary shares. *J. Air Transp. Mgmt.*, 9(6):361–369, 2003.
5. G. Dobson and P. Lederer. Airline scheduling and routing in a hub-and-spoke system. *Transportation Science*, 27(3):281–297, August 1993.
6. M. Etschmaier and D. Mathaisel. Airline scheduling: An overview. *Transportation Science*, 19(2):127–138, May 1985.
7. R. Gopalan and K. Talluri. Mathematical models in airline schedule planning: A survey. *Annals of Operations Research*, 76:155–185, 1998.
8. C. Hane et al. The fleet assignment problem: Solving a large-scale integer program. *Mathematical Programming*, 70:211–232, 1995.
9. M. Lohatepanont and C. Barnhart. Airline schedule planning: Integrated models and algorithms. *Transportation Science*, 38(1):19–32, February 2004.
10. D. Mathaisel. Decision support for airline schedule planning. *Journal of Combinatorial Optimization*, 1:251–275, 1997.
11. K. Proussaloglou and F. Koppelman. The choice of air carrier, flight, and fare class. *Journal of Air Transport Management*, 5(4):193–202, 1999.
12. F. Soumis, J. A. Ferland, and J.-M. Rousseau. A model for large scale aircraft routing and scheduling problems. *Transportation Res.*, 14B:191–201, 1980.
13. K. Talluri and G. van Ryzin. *The theory and practice of Revenue Management*. Kluwer Academic Publishers, 1st edition, 2004.

‘T’ for Tabu and Time Dependent Travel Time

Johan W. Joubert

Department of Industrial and Systems Engineering, University of Pretoria, South Africa

johan.joubert@up.ac.za

Summary. Even in its most basic form, the Vehicle Routing Problem and its variants are notoriously hard to solve. More often artificially intelligent algorithms are employed to provide near-optimal solutions. To be classified as “intelligent”, however, a solution strategy should be able to first analyze the environment in which the problem occurs, then solve it, and afterwards reflect on the solution process so as to improve future decision-making. Although reference is made to the entire context of the intelligence research project, this paper reports on the Tabu Search algorithm designed that catered for a problem with multiple soft time windows, a heterogeneous fleet, double scheduling, and time dependent travel time. An adaptive memory procedure was employed, initially populated with good initial feasible solutions, and the algorithm was tested on 60 problems based on established benchmark sets. The complex variant of the Vehicle Routing Problem required between 670 and 4762 seconds on a standard laptop computer, which is considered to be reasonable in the proposed application, and was consistent between different runs with an absolute mean deviation of 3.6%. The contribution is significant as it provides an algorithm that efficiently addresses a complex and practical application of the Vehicle Routing Problem. The algorithm can easily be extended to make use of multiple processors so as to reduce computational time.

1 Introduction

The Vehicle Routing Problem (VRP) has its aim to assign customer deliveries to vehicles, and determining the visiting order of each vehicle. The VRP is a well-researched problem in Operations Research literature. The problem is typically solved in engineering and logistics situations where the distribution of products are concerned – the main objective being the minimization of an objective function, typically distribution cost for individual carriers. In its basic form, the VRP can be described as the problem of assigning optimal delivery or collection routes from a depot to a number of geographically distributed customers, subject to problem-specific constraints. The most basic version of the VRP can be defined with $G = (V, E)$ being a directed

graph where $V = \{v_0, v_1, \dots, v_n\}$ is a set of vertices with v_0 representing the depot where m identical vehicles, each with capacity Q , are located. The remaining vertices, denoted by $V \setminus \{v_0\}$, represents customers each having a non-negative demand q_i and a non-negative service time s_i [8]. The edge set connecting the vertices is given by $E = \{(v_i, v_j) | v_i, v_j \in V, i \neq j\}$. A distance matrix $C = \{c_{ij}\}$ is defined on E . In some contexts, c_{ij} can be interpreted as travel cost or travel distance. Hence, the terms distance, travel cost, and travel time are used interchangeably. The VRP consists of designing a set of m vehicle routes having a minimum total length such that

- each route starts and ends at the depot,
- each remaining vertex ($V \setminus \{v_0\}$) is visited exactly once by one vehicle,
- the total demand of a route does not exceed Q , and
- the total duration (including service and travel time) of a route does not exceed a preset limit L .

The VRP is a hard combinatorial (NP-hard) optimization problem for which several exact and approximate solution methods have been proposed (see Laporte [5] for a review). The area of application is wide, and specific variants of the VRP transform the basic problem to conform to application specific requirements. Additional constraints are introduced to adapt to business-specific requirements, and yields a hard combinatorial problem by restriction. The more complex a problem become, the more favorable the choice of approximate as opposed to exact algorithms due to the computational burden.

Metaheuristics are master strategies which use intelligent decision making techniques to guide algorithms to find global optimal solutions by temporarily allowing moves in the neighborhood that result in solutions with inferior objective function values compared to the incumbent solution. The majority of metaheuristics follow a similar process, although their specific naming conventions, analogies and detailed routines may vary.

Initialization – The process of finding an initial solution. Some metaheuristics, such as the *Genetic Algorithm* performs well with randomly generated initial solutions that need not be feasible, while the *Tabu Search* is highly sensitive to the quality of the initial solution.

Diversification – The mechanism that ensures that the underlying heuristics are searching a diversified neighborhood, and thus not getting trapped within local optima.

Intensification – A mechanism that ensures the heuristic starts *zooming in* towards a single solution. The most promising neighborhoods are identified and those areas of the solution space are searched more thoroughly.

The diversification and intensification can repeat indefinitely, and hence requires a stopping criteria to terminate the metaheuristic [8]. The longer the metaheuristic runs, the higher the probability of converging to the global optimum. The interested reader is referred to [7] for a recent and comprehensive review of various solution strategies. In a comparison of heuristics

and metaheuristics, [8] identifies Tabu Search as a dominant improvement heuristic with the certainty of achieving at least a *local* optimum. The Tabu Search implemented in [2] indicated that the model provided substantial improvements over a model based on fixed travel times. Recent developments include drastically reducing the size of the search neighborhood, so called *granular* neighborhoods [6]. Results obtained when using promising moves, as proposed by granular neighborhoods, yielded good solutions within short computing times.

In this paper the initial solution heuristic provided feasible solutions that already catered for multiple soft time windows, a heterogeneous fleet, and double scheduling, all while taking time dependent travel times into account. Multiple initial solutions were passed to the Tabu Search algorithm that stored the most promising tours (set of routes assigned to a single vehicle) in an adaptive memory structure. The algorithm iteratively generated a new solution from the adaptive memory. A self-regulating mechanism was introduced to select between two exchange operators. Once a new promising solution was completed, it was decomposed into tours, which in turn was stored in the memory structure. A predefined iteration limit was used to terminate the algorithm.

The Tabu Search algorithm provided good results to adapted test problems and performed consistently between runs with an absolute mean deviation among 240 runs of only 3.6%.

Introducing a self-regulating mechanism enabled the algorithm to adapt to its environment. The contribution of this paper is therefor not only an algorithm that solves a complex and practical variant of the VRP, but also an algorithm that will be useful to developers of routing solutions in a variety of problem environments challenged with selecting a single appropriate algorithm.

The paper is structured as follows: Section 2 provides the structure of both the initial solution and the main Tabu Search algorithm. Results of the algorithm is given in Section 3. The paper is concluded in Section 4 with a brief discussion and a few final remarks on future implementation.

2 The Tabu Search Algorithm

An exact travel-time matrix was established that adheres to the triangular inequality. To avoid the excessive computational burden when calculating and checking the matrix, time windows and time window compatibility was used to eliminate obvious infeasible origin-destination combinations [4]. The initial solution algorithm used in [4] only provided a single solution. As I , preferably *different*, initial solutions were required, a random node $I_{i \in \{1,2,\dots,I\}}^* \in \{1,2,\dots,n\}$, with n the total number of customer nodes, was identified and removed from the problem set P . The remaining nodes in $P' = P \setminus \{I_i^*\}$ were used to create an initial solution using the improved initial

solution algorithm proposed in [4]. After the nodes in P' have been routed, the identified node I_i^* was reinserted into the first feasible position. The result was a set of initial solutions $\mathbf{s} = \{s_1, s_2, \dots, s_I\}$. Each initial solution was decomposed and its tours were stored in the adaptive memory, associated with it the objective function value of the initial solution from which the tour originates.

A partially constructed tour was created through iteratively selecting tours from the adaptive memory, and removing all tours from the adaptive memory that shared nodes with the selected tour. The probability of selecting any tour was based on the objective function associated with the tour, which in turn was taken from the solution from which the tour originates. It is noted in [1] that the use of probabilities, based on past performance, as an underlying measure of randomization yields efficient and effective means of diversification. The better a solution, the higher the probability of selecting a tour from that solution. Once a tour was selected from the adaptive memory, all tours sharing nodes with the selected tour were removed from memory. Removing tours from the adaptive memory ensured each node was represented only once in the partially constructed tour. The selection of tours from the adaptive memory, and the removal of tours with common nodes, was repeated until no more tours remained in the adaptive memory. As not all nodes were represented, the partially constructed tour denoted by s , was completed by inserting the remaining unrouted nodes into feasible positions of s . The resulting tour, denoted by s^* , was achieved through either identifying positions on a current route, creating a new route on a current tour, or creating a new tour with its associated vehicle.

Two exchange operators were considered. The first operator removes a randomly selected node from one tour and inserts the node into the best possible position in another tour that has the same vehicle type. The second operator also removes a randomly selected node from an origin tour, but selects the best insertion position for the node on a tour having a different vehicle type than the origin tour.

Initially the probability of selecting either of the operators was equal. A frequency parameter, ζ , ensures that every ζ iterations *both* operators were used to create perturbations. The probability of the operator producing the best solution was then increased relative to its current probability.

3 Results

To accommodate not only multiple time windows, but also a heterogeneous fleet, test data was created according to [4]. Results for two randomly selected problems are provided in Fig. 1 where both the current objective function value, and the incumbent is illustrated. The convergence of the objective function, as well as the continued diversification is prominent.

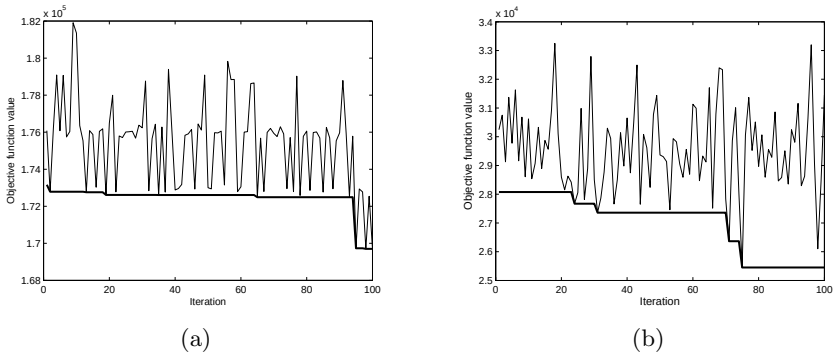


Fig. 1. Random selection of result graphs using the Tabu Search, illustrating both convergence of the objective function, and the diversified solution space considered. The thin line indicates the current objective function value for the associated iteration, while the thicker black line indicates the incumbent objective function value

The objective function constitutes total travel cost, a weighted lateness penalty resulting from soft customer time windows being violated, and total fleet cost representing the heterogeneous fleet.

Due to the random component inherent in the algorithm, each problem was solved four independent times. A total of 240 problem instances were solved with an absolute mean deviation of approximately 3.8% indicating good repeatability of the algorithm's execution. Due to limited space in this paper, actual results have been omitted, and can be requested from the author. Results indicate for each of the 60 problems the incumbent solution, the associated fleet assignment, as well as the computational time required for each of the four independent runs.

4 Discussion and Conclusion

For the first iteration of the execution the incumbent solution was the best initial solution prior to being decomposed and added to the adaptive memory. The current solution of the first iteration, however, was the first solution created from the adaptive memory, often resulting in a solution worse than the incumbent. This is illustrated in both cases presented in Fig. 1.

The initial solution algorithm used to initially populate the adaptive memory produced high quality solutions that was never, over 240 problems solved, improved by more than 10% over 100 iterations. This not only confirms the quality of the initial solution algorithm, but also indicates the Tabu Search's sensitivity for a good quality initial solution.

This paper provided an algorithm that solved a complex and practical variant of the notoriously hard Vehicle Routing Problem with test problems of

100 customers between 670 and 4762 seconds on a standard single-processors Pentium CentrinoTM laptop. Although time-dependent travel time data is becoming more readily available, more work is required to widely implement the proposed research. The algorithm provided in this paper has been incorporated along with other solution algorithms in an intelligent agent suggested by [3]. The agent is intended to be implemented in areas where the structure of customer data is unknown. Once implemented, the agent perceives the environment through fuzzy cluster analysis; predicts the best solution algorithm to be used; and periodically reflects on its predictions for improved future decision-making using an artificial neural network.

References

1. Glover, F. (1990). Tabu search — part II. *ORSA Journal on Computing*, 2(1):4–32.
2. Ichoua, S., Gendreau, M., and Potvin, J.-Y. (2003). Vehicle dispatching with time-dependent travel times. *European Journal of Operational Research*, 144(2):379–396.
3. Joubert, J. W. (2006). Vehicle routing: less “artificial”, more “intelligence”. In Brebbia, C. A. and Dolezel, V., editors, *Urban Transport and the Environment in the 21st Century*, volume 89 of *WIT Transactions on The Built Environment*, pages 245–256, UK. WIT Press.
4. Joubert, J. W. and Claasen, S. J. (2006). A sequential insertion heuristic for the initial solution of a constrained vehicle routing problem. *ORiON*, 22(1):105–116.
5. Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59(3):345–358.
6. Toth, P. and Vigo, D. (2003). The granular tabu search and its application to the vehicle-routing problem. *INFORMS Journal on Computing*, 15(4):333–346.
7. Toth, P., Vigo, D., and (editors) (2002). *The Vehicle Routing Problem*. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.
8. Van Breedam, A. (2001). Comparing descent heuristics and metaheuristics for the vehicle routing problem. *Computers & Operations Research*, 28(4):289–315.

Integrated Operational Transportation Planning in Theory and Practice

Herbert Kopfer and Marta Anna Krajewska

Chair of Logistics, University of Bremen, Germany

makr@logistik.uni-bremen.de

kopfer@logistik.uni-bremen.de

1 Introduction

Outsourcing (subcontraction) makes use of external entities as an alternative to the usage of own resources (self-fulfillment). The arising 'make-or-buy' decision combining the two clusters of self-fulfillment and subcontraction evolves in a complex reference analysis among items involved [13].

In particular, considering the integrated operational transportation planning, the 'make-or-buy' decision also applies to the logistic branch. In practice freight forwarders face great demand fluctuations regarding transportation volume [1], while their own vehicle fleet is associated with high maintenance costs. The remedy is to reduce capacity of the own vehicle fleet far under the varying total demand limit. Instead, additional outside carriers are involved in order to gain enough transportation resources for covering the demand. Together with engaging an external forwarder, the methods of payment for its service (types of subcontraction) are defined. Furthermore, planning methods for three arising tasks are established: assignment to the execution mode and cost optimisation within each mode. In the following sections we present how these decisions can vary in theory and practice. Section 2 investigates the frames for modelling the integrated operational transportation planning problem. Section 3 introduces different theoretical and practical types of subcontraction. Section 4 compares theoretical with practical planning methods.

2 Modelling the Integrated Operational Transportation Planning Problem

There are only a few theoretical proposals for modelling an integrated operational transportation planning problem. They all assume that an order is a pickup-and-delivery request describing a single transportation demand, which results in a direct transportation process of carrying a less-than-truckload

packet. The aim is to minimize the total costs of request fulfillment, which is the sum of costs in the self-fulfillment and subcontraction cluster. The research for self-fulfillment is focused on the widely known pick-up-and-delivery-problem-with-time-window-constraints (compare [2]). The round tours are constructed, whose length results from routing and scheduling of vehicles. The homogeneous ([8] [11]) alternatively heterogeneous ([1] [12] [9] [3]) vehicle fleet is stationed in one depot (except for [3]) and each vehicle has the same predefined vehicle capacity. The tariff rate per distance unit, possibly time unit ([8] [3]), and additionally fixed costs per vehicle ([1]), provide a basis for cost calculation. Such modelling assumptions for the self-fulfillment cluster correspond to the reality in many freight forwarding companies. In contrast to self-fulfillment, there exists no unique way to calculate the execution costs of requests in the subcontraction cluster.

3 Different Types of Subcontraction

The models of integrated operational transportation planning proposed so far incorporate different types of subcontraction. Further subcontraction models have been introduced in literature as separate problems or have arisen from our practical analysis.

Firstly, the theoretical approaches are the subject of consideration. The easiest method among those approaches simply shifts requests representing a sale of a single request independently from all the other requests to an external freight carrier ([1]). The requests are then forwarded on uniform conditions, based on a fixed fee per distance unit, for the distance between the pickup and delivery location.

A more complicated method assumes complete tours to be shifted to subcontractors ([9] [12]). The applied method of cost calculation corresponds to the cost calculation for the self-fulfillment cluster. The difference between the two approaches consists in the used terms of hiring vehicles. In [12] it is assumed that the vehicles of the external freight forwarder are leased in the short term if they are needed. Thus, a variable number of vehicles is possessed and only an exploited part of maintenance costs (percentage of working times of vehicles and drivers) is covered. In [9] a part of vehicles is rented permanently, i.e., they cannot be returned, while the other part is leased at short notice and sent home if the number of requests decreases. As the maintenance costs for the vehicles from the first group have to be fully covered, the minimal number of hired vehicles is aimed by the usage of punishment costs.

Requests can be forwarded independently of each other, where the costs of subcontracting the requests are adjusted to the effort in case of self-fulfillment ([3] [11]). Hence, the freight cost calculation results from isolated price assessment for each request on the basis of a fixed tariff, multiplied with the adjustment parameter defined for different criteria (distance, weight). In the approach of [11] an additional round route for an artificial vehicle is constructed. All the

requests that are assigned to the subcontraction cluster are planned into this round route. The adjustment parameter is defined as a quotient of the driven distance of the artificial route and the sum of distances between the pickup and the delivery locations for all the requests in the subcontraction cluster. Analogically, [3] defines the adjustment parameter as the quotient of the volume transported by the particular vehicle from the own fleet and the average vehicle capacity of the fleet.

The most complex theoretical method assumed for subcontraction is called freight flow consolidation ([8]). The problem belongs to the multi-commodity network flow problems: the less-than-truckload requests are bundled and a cost minimal flow through a given transportation network is sought. The construction of a set of spanning trees, containing the subsets of requests is aimed. The freight flow consolidation problem itself is a combined non-linear flow and assignment NP-hard optimisation problem. There are no fixed costs connected with such fulfillment, only the variable costs which depend on the distance and loading weight ([5] [6]). Typically, costs are degressive with respect to increasing weight of the bundled goods, which results in a non-linear freight function.

Freight flow consolidation in presence of time windows has been presented by [10] within the determination of the so-called origin/destination paths. It has not been introduced to the integrated operational transportation planning problem yet.

Secondly, subcontraction possibilities that result from our analysis of practical approaches in freight forwarding companies are considered. The analysis has shown that for the subcontraction of less-than-truckload request bundles between independent partners the freight flow consolidation method (without time windows constraints) remains of highest practical relevance. There usually exist fixed tariff tables under non-linear consideration of distance, weight of a bundle and the type of goods that should be transported. The price for customers is quoted on the basis of such tables, although it can be modified dependent on the driven direction.

In case of full-truckload request bundles two methods of subcontraction come into consideration. As already mentioned, complete tours can be shifted to subcontractors on a fixed tariff basis which is mostly dependent on distance of the round route to be driven. There are no fixed costs connected with such usage of foreign vehicles, but the tariff rate for the variable costs is higher in comparison to the own vehicle fleet as it covers a part of maintenance costs (compare [12]). The second possibility consists in payment for subcontractors on a daily basis. In this case an external freight forwarder gets a daily flatrate and has to fulfill all the received requests up to an agreed distance and time limit. Costs arising within both subcontraction possibilities are compared with the costs for self-fulfillment in figure 1.

All theoretical approaches we have found in literature involve only one type of subcontraction. In practice single freight forwarders apply many forms of payment for outsourcing that are used simultaneously.

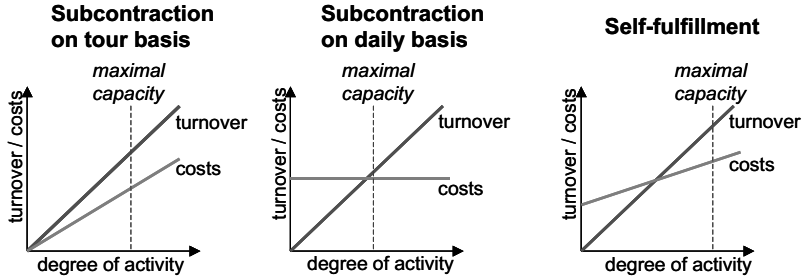


Fig. 1. Comparison of costs for different types of subcontraction and self-fulfillment [4]

4 Planning Methods

Different methods of combining self-fulfillment and subcontraction clusters in solving techniques result in three main types of integration: hierarchical, semi-hierarchical and simultaneous integration can be found. The types of integration are outlined in figure 2.

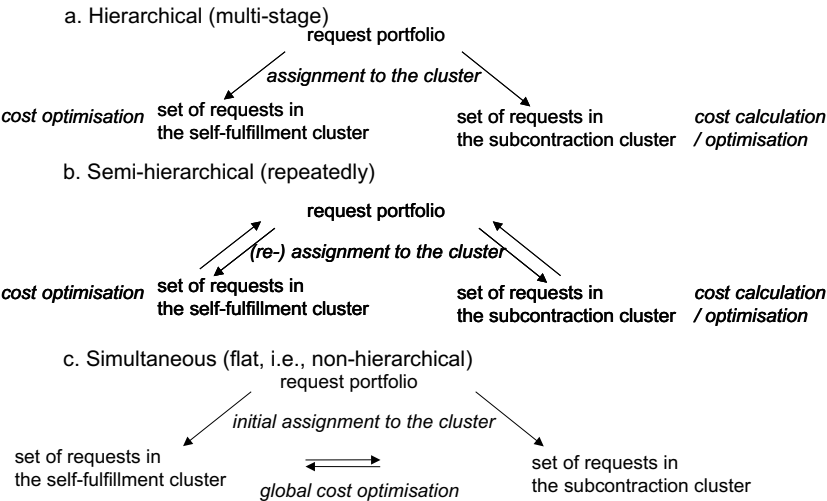


Fig. 2. Different types of cluster integration

In case of hierarchical planning (multi-stage planning), the request portfolio is splitted into two subportfolios that are assigned to the clusters. After that the cost optimisation takes place inside the clusters. Such an approach is presented by [1]. The semi-hierarchical planning procedure, introduced by [8],

runs repeatedly. Request bundles are assigned to the self-fulfillment or subcontraction clusters. Next, the optimisation procedures run in the single clusters. Afterwards, the requests are reassigned to the clusters. As the complex optimisation tasks in both clusters cause high time-consumption for the accurate cost calculation, the semi-hierarchical planning allows changes in the division of the request portfolio into the clusters only on the basis of cost assessments. In the simultaneous (flat, i.e., non-hierarchical) approach ([11] [3] [9] [12]) requests are initially assigned to one of both clusters and then the global cost optimisation procedure takes place. There exists a complete implementation plan for all requests in each iteration. The modification of such a plan for the next iteration runs on a global level: in order to get the next modified plan single requests are shifted inside as well as between the clusters.

As the integrated operational transportation planning problem is NP-hard, different complex solution methods have been applied to test instances. However, the theoretical frames have not been put into practice yet, as the practical complexity of the problem is much higher. There usually are different types of subcontraction that should be combined at a single blow. Till now, no theoretical approach proposes an integrated solution where several types of subcontraction methods appear (see sect. 3). Moreover, the problem has to be handled dynamically as there is a constant flow of request orders to the freight forwarder during the planning period. Thus, in the freight forwarding companies, where both self-fulfillment and subcontraction are involved, requests are planned manually. Planning is made hierarchically (compare [4]). In the first place the requests with the highest contribution margin are planned into the self-fulfillment cluster. Here, schedulers can be supported by software that optimises the subproblem of building round routes. Then the other types of subcontraction are considered, also in a hierarchical order. The advantages of simultaneous planning are lost.

5 Conclusions

The integrated operational transportation planning problem reflects the practical situation of today's planning in freight forwarding companies. However, our analysis of existing optimisation software systems for freight forwarders has shown that the problem is underestimated. Neither the subcontraction possibilities nor the advanced simultaneous optimisation methods beyond the user interface are involved at all. Such situation is caused on one hand by a still insufficient theoretical background in the existing approaches (see sect. 2) and, on the other hand, by a missing link between science and economy. Yet, logistic market gets more and more competitive, and the possibilities to decrease costs are almost exhausted, freight forwarders slowly discover that they cannot work efficiently without structural changes which include establishing fully automatic and efficient planning. Planning systems for the integrated operational transportation planning, considering several subcontraction pos-

sibilities and simultaneous solving methods producing high quality fulfillment schedules are becoming inevitable.

References

1. Chu C (2005) A heuristic algorithm for the truckload and less than truckload problem. *European Journal of Operations Research* 165: 657–667
2. Dumas Y, Desrosiers J, Soumis F (1997) The pickup and delivery problem with time windows. *European Journal of Operations Research* 54: 7–22
3. Greb T (1998) Interaktive Tourenplanung mit Tabu Search. PhD Thesis, University of Bremen, Bremen
4. Jurczyk A, Kopfer H, Krajewska M (2006) Speditionelle Auftragsdisposition eines mittelständischen Transportunternehmens. *Internationales Verkehrswesen* 6: 275–279
5. Kopfer H (1990) Der Entwurf und die Realisierung eines A*-Verfahrens zur Lösung des Frachtoptimierungsproblems. *OR Spectrum* 12: 207–218
6. Kopfer H (1992) Konzepte genetischer Algorithmen und ihre Anwendung auf das Frachtoptimierungsproblem im gewerblichen Güterfernverkehr. *OR Spectrum* 14: 137–147
7. Lau H, Liang Z (2001) Pickup and delivery problem with time windows: algorithms and test generation. *Proceedings IEEE Conference on Tools with Artificial Intelligence*: 222–340
8. Pankratz G (2002) Speditionelle Transportdisposition. PhD Thesis, University of Hagen, Wiesbaden
9. Savelsbergh M, Sol M (1998) DRIVE: Dynamic routing of independent vehicles. *Operations Research* 46: 474–490
10. Schönberger J, Kopfer H (2004) Freight flow consolidation in presence of time windows. *Operations Research Proceedings 2004*: 184–192
11. Schönberger H (2005) *Operational Freight Carrier Planning*. Springer, Berlin Heidelberg New York
12. Stumpf P (1998) Tourenplanung im speditionellen Güterfernverkehr. Schriftreihe der Gesellschaft für Verkehrsbetriebswirtschaft und Logistik, Nürnberg
13. Wellenhofer-Klein M (1999) *Zulieferverträge im Privat- und Wirtschaftsrecht*. Beck Verlag, München

Managerial Accounting and Auditing

Investment Incentives from Goal-Incongruent Performance Measures: Experimental Evidence

Markus C. Arnold, Robert M. Gillenkirch, and Susanne A. Welker

Department of Finance, Accounting and Taxes, University of Göttingen, Germany
markus.arnold@wiwi.uni-goettingen.de (Corresponding author)
rgillen@uni-goettingen.de
swelker@uni-goettingen.de

1 Introduction

We analyze investment incentives of goal incongruent performance measures in an experiment in which ‘managers’ make one-period investment decisions and ‘owners’ predict these decisions. Three alternative performance measures are considered: earnings, ROI, and residual income. These measures serve as archetypes for a wide variety of measures used in practice. Standard theoretical predictions with respect to the investment incentives of earnings, ROI, and residual income are well documented, and they have become part of the management accounting education (e.g., [5]). They illustrate one of the basic principles of management accounting: ‘you get what you pay for’. As is well known, neither earnings nor ROI is a goal congruent performance measure, whereas residual income is goal congruent due to its conservation property.

Empirical (field) studies (e.g. [7]; [4]; [2]) have examined changes in firms’ operating, investment, and financing decisions, and in stock price performance following the adoption of residual income measures. Overall, the studies support the theoretical predictions, although some of the reported effects are not as strong as predicted.

In an organization, every decision is embedded in a social context that influences an individual’s behavior and thus has an impact on the effectiveness of management accounting systems; see [6]. In contrast to standard theoretical predictions, experimental economics has consistently shown that people do not always maximize their individual payoffs (e.g., [3]). Thus, we cannot take for granted the motivational effects of different performance measures derived from standard theory.

In our experiment, we find that, although managers overinvest (underinvest) on average when they are evaluated at earnings (ROI), owners do not get what they pay for from the majority of managers when using these two performance measures. Nevertheless, owners overestimate these measures’ effectiveness in inducing individually rational behavior. When confronted with

managers’ decisions in early rounds, owners consistently revise their beliefs. However, their overall estimation quality remains relatively low.

2 Experimental Design

The experiment consisted of three treatments referring to the performance measures to be examined, earnings (E), Return on Investment (ROI), and residual income (RI), where $E = r(x) - x$, $ROI = E/x$, $RI = E - \kappa x$. x is the amount invested, $r(x)$ is the cash return, and κ is the cost of capital. Participants in the experiment were assigned either the role of a ‘manager’ or of an ‘owner’ in one of the three treatments. Managers received a share of the respective measure, and owners received RI , which equals the project’s terminal value, minus the manager’s payoff. The shares were set such that managers would receive comparable compensations from the experiment if they always maximized their own payoff. Managers and owners were paired randomly and anonymously in every round; owners were passive in that they could not influence managers’ decisions.

A manager’s task consisted in making investment decisions in 12 rounds. In each round, a manager received a decision sheet that displayed 12 amounts x that could be invested, the corresponding cash returns $r(x)$, the ‘net surplus’ (RI), and the manager’s and the owner’s payoffs. Table 1 presents the managers’ decision problem in round 1. In the table, a treatment-specific category index k is added. In order to simplify comparisons we normalize the category that maximizes the manager’s payoff to $k = 0$ and term it as the ‘individually rational category’. The value maximizing category is $k = -4$ in the earnings treatment, $k = 4$ in the ROI treatment, and $k = 0$ in the residual income treatment. Thus, an individually rational choice of $k = 0$ in the earnings/ROI

Table 1. Investment project and owner and manager payoffs in round 1.

Inv. amount	35	42	56	70	84	98	112	126	140	154	160	164
Cash return	55.15	74.30	95.73	116.43	136.63	156.46	173.12	189.77	206.43	223.08	224.70	225.20
category k	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2
Earnings	20.15	32.30	39.73	46.43	52.63	58.46	61.12	63.77	66.43	69.08	64.70	61.20
Man. payoff	3.93	6.30	7.75	9.05	10.26	11.40	11.92	12.44	12.95	13.47	12.62	11.93
Own. payoff	5.72	13.40	15.18	16.38	17.17	17.66	15.60	13.53	11.48	9.41	4.08	0.07
category k	-1	0	1	2	3	4	5	6	7	8	9	10
ROI	0.576	0.769	0.710	0.663	0.627	0.597	0.546	0.506	0.475	0.449	0.404	0.373
Man. payoff	10.36	13.84	12.77	11.94	11.28	10.74	9.82	9.11	8.54	8.07	7.28	6.72
Own. payoff	-0.71	5.86	10.16	13.49	16.15	18.32	17.70	16.86	15.89	14.81	9.42	5.28
category k	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Res. inc.	9.65	19.70	22.93	25.43	27.43	29.06	27.52	25.97	24.43	22.88	16.70	12.00
Man. payoff	4.83	9.85	11.47	12.72	13.72	14.53	13.76	12.99	12.22	11.44	8.35	6.00
Own. payoff	4.83	9.85	11.47	12.72	13.72	14.53	13.76	12.99	12.22	11.44	8.35	6.00

treatment implies over-/underinvestment. The 12 decision problems differed with respect to the concrete values for x and $r(x)$, but the value maximizing category was always $k = -4/4$ in the earnings/ROI treatment.

An owner's task consisted in predicting the decisions of all managers' in two treatments – her own and a second treatment – in all 12 rounds. This controls for effects of 'wishful thinking'; however, we did not find any such effects. Managers were not informed that owners would estimate their decisions. The estimation consisted in predicting, for every possible investment amount, the number of managers who would choose this amount. For this task, owners were compensated according to the accuracy of their estimations which was measured by the following absolute error:

$$ae_{jt} = \sum_k |n_t^k - \hat{n}_{jt}^k| + \left| N - \sum_k \hat{n}_{jt}^k \right| \quad (1)$$

where n_t^k is the number of *actual* choices of the investment amount in category k and $\hat{n}_{jt}^k$ is the number of choices owner j *estimates* to be in category k in round t . N is the total number of decisions to be estimated.¹ After round 4 and round 8, owners were informed about managers' actual decisions in all preceding rounds. Thus, when making their estimations in rounds 1-4, owners had no information about managers' actual decision behavior, whereas in rounds 5-8 and 9-12, they became increasingly 'experienced'.

A manager's second task was to estimate the decisions in his treatment in three selected rounds.² Owners' and managers' total compensations from the experiment consisted of the payoff from a manager's decision in a randomly selected round plus a payoff according to the accuracy of their estimations, as measured by (1). On average, participants earned a total of 17.74 EUR in the experiment, with a minimum of 8.00 EUR and a maximum of 24.64 EUR.

The experiment was run at the University of Göttingen in two sessions. In total, 130 undergraduate students participated at the experiment, 66 in the manager group and 64 in the owner group. The organization of the experiment guaranteed anonymous decision making.

3 Results

Managers' decisions : Figure 1 displays the average number of manager decisions per round in the three treatments in all 12 rounds. The figure shows that, although the individually rational category $k = 0$ is the managers' modal choice, it is chosen in 37.50% (31.06%) of the cases in the earnings (the ROI) treatment only, which is far below the standard theoretical prediction. In these treatments, managers tend towards the owners' optimum and the value maxi-

¹ The second term in (1) is a penalty term which was added to ensure that participants could not improve their estimation quality by estimating more or less than the required number of decisions.

² The results with respect to these estimation are not reported; see [1].

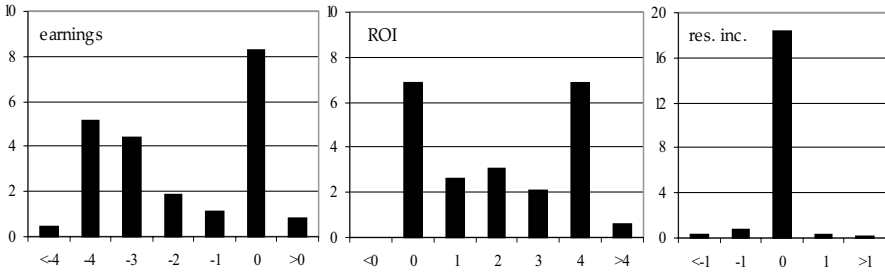


Fig. 1. Average number of managers' decisions per round in all 12 rounds. Earnings, ROI and residual income treatment.

mizing investment amount in the majority of the cases. In the ROI treatment, the number of value maximizing investment choices is even as large as the number of individually rational decisions. Overall, in 58.71% of the cases in the earnings treatment and in 68.94% of the cases in the ROI treatment, owners receive a higher payoff than under standard theoretical predictions. The corresponding actual mean owner payoff exceeds the expected mean owner payoff according to standard theory ($k = 0$ in all rounds) by 62.98% in the earnings treatment and by 114.79% in the ROI treatment. Nevertheless, Figure 1 also reveals that managers do not choose the value maximizing category in the majority of the cases, either, but overinvest on average in the earnings treatment and underinvest on average in the ROI treatment. Finally, the figure confirms that, in the residual income treatment as a control treatment, more than 92% of the decisions were according to the standard theoretical prediction ($k = 0$). Mean owner payoffs in this treatment were significantly larger than in the two other treatments. Thus, our results provide evidence for agency conflicts due to the goal incongruence of a performance measure, although these agency problems are smaller than predicted.

Owners' estimations : First, we analyze owners' estimations in rounds 1-4. Figure 2 contrasts the average number of decisions per round estimated by all owners for every category in rounds 1-4 to the average actual number in these categories. Owners' estimates in the earnings and ROI treatments strongly deviate from the standard theoretical prediction as owners predict only 13.1 (59.53%) individually rational decisions on average in the earnings treatment, and only 13.7 (62.40%) individually rational decisions in the ROI treatment.³ Though, Figure 2 also shows that, when compared to the actual decisions in the earnings and ROI treatments, owners still largely overestimate the number of individually rational decisions. For earnings, the estimated number of decisions in category $k = 0$ is nearly 60% larger than the actual number,

³ In contrast, in the residual income treatment, owners by and large expect to get what they pay for, as nearly 90% of the decisions are estimated to be individually rational, and $|k| \leq 1$ in over 95% of the cases.

and for the ROI, it is even more than twice as large as the actual number of decisions in this category. Correspondingly, owners underestimate especially the number of value maximizing investment decisions in both treatments.

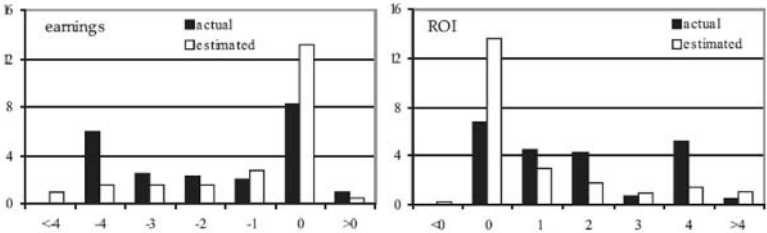


Fig. 2. Estimated vs. actual decisions. Earnings and ROI treatment, rounds 1-4.

This result suggests that unexperienced owners overestimate the effectiveness of performance measures in inducing individually rational behavior.

Now consider the owners’ estimations in rounds 5-12 after learning managers’ actual decisions in the preceding rounds. Our experimental design allows testing the owners’ estimation accuracy with respect to the complete distribution of managers’ decisions as well as with respect to the mean pay-offs from these decisions. In the following, we report results with respect to the estimation of distributions only; for a detailed analysis of owners’ estimations, see [1]. Table 2 displays the evolution of owners’ absolute estimation errors.

Table 2. Evolution of the average owner estimation errors.

	abs. estimation error <i>ae</i>			significance: p-values		
	1-4	5-8	9-12	1-4/5-8	5-8/9-12	1-4/9-12
Earnings	22.82	19.88	17.13	0.004***	< 0.001***	< 0.001***
ROI	25.63	18.71	15.34	< 0.001***	< 0.001***	< 0.001***
Residual income	6.21	2.19	4.15	< 0.001***	< 0.001***	0.489

In both the earnings and the ROI treatment, owners’ estimations improve continuously over time. As the table shows, the improvements are highly significant.⁴ Obviously, the information provided after rounds 4 and 8 makes owners adjust their estimations consistently. However, owners’ estimation errors in these two treatments are largely positive even in rounds 9-12, and they are by far larger than the corresponding errors in the residual income treatment. In order to assess the magnitude of estimation errors in the earnings and the ROI treatment, we compared the actual owners’ estimations to stylized estimations. In rounds 9-12, a stationary estimation, which estimates

⁴ Units of observation: individual averages across rounds 1-4, 5-8, 9-12. Wilcoxon rank-sum tests. ***, ** and * denote significance at the 1%, 5% and 10% level.

decisions in future rounds by simply extrapolating past decisions, outperforms 90% of the owners who estimate earnings managers' decisions, and 85% of the owners estimating ROI managers' decisions. Even a random distribution of estimations over the five categories 0 to -4 (0 to 4) in the earnings (ROI) treatment outperforms more than 50% (more than 60%) of the respective owners' estimations. We interpret this as strong evidence against the owners' ability to correctly predict the distribution of decisions.

4 Conclusion

Our study addresses the investment incentive effects of two goal incongruent performance measures, earnings and ROI. We analyze these effects in an experiment with a control treatment referring to the goal congruent residual income measure. We interpret our findings as evidence for the existence of agency conflicts in the context of managerial investment decisions if goal incongruent performance measures are used. However, the magnitude of these conflicts are smaller than predicted by standard theory. Second, we also interpret the results of our experiment as evidence against the use of goal incongruent performance measures in compensation practice. This conclusion is not only derived from the observed under- and overinvestment in the ROI and the earnings treatment but receives particularly strong support from our findings on owners' estimation quality. Obviously, a goal congruent measure not only induces better decisions, but also largely reduces the complexity of forecasting these decisions.

References

1. Arnold, M. C., R. M. Gillenkirch, and S. A. Welker (2006) Do You Get What You Pay For? An Experimental Analysis of Managers' Decisions and Owners' Expectations. Working Paper, University of Göttingen.
2. Balachandran, S. V. (2006) How Does Residual Income Affect Investment? The Role of Prior Performance Measures. *Management Science* 52: 383–394.
3. Fehr, E. and A. Falk (2002) Joseph Schumpeter Lecture - Psychological foundations of incentives. *European Economic Review* 46: 687–724.
4. Hogan, C. E. and C. M. Lewis (2005) Long-Run Investment Decisions, Operating Performance, and Shareholder Value Creation of Firms Adopting Compensation Plans Based on Economic Profits. *Journal of Financial and Quantitative Analysis* 40: 721–745.
5. Solomons, D. (1985) *Divisional Performance: Measurement and Control*. Illinois: Wiener Publishing, 2 edition.
6. Sprinkle, G. (2003) Perspectives on Experimental Research in Managerial Accounting. *Accounting, Organizations and Society* 28: 287–318
7. Wallace, J. S. (1997) Adopting residual income-based compensation plans: Do you get what you pay for?. *Journal of Accounting and Economics* 24: 275–300

Metaheuristics and Decision Support Systems

Modelling Qualitative Information in a Management Simulation Game

Volker Nissen and Giorgi Ananidze

FG Wirtschaftsinformatik für Dienstleistungen, TU Ilmenau, Germany
volker.nissen@tu-ilmenau.de

1 Management Simulation Games and the Importance of Qualitative Information

Management simulation allows students to apply their theoretical knowledge in a safe, feedback-learning context. The quality of the simulation model is important for the achievable learning effect. A model always simplifies reality, but should still strive for a realistic implementation of key factors and relationships of the modeled situation. Vague information and qualitative relationships have a great significance in practical business. In management games, however, such qualitative aspects are often excluded from the simulation as it is difficult to model them adequately [1]. Concepts like "competence", "image" and "credit rating" are meaningful to business, but difficult to grasp with conventional simulation modeling. In our contribution an approach based on fuzzy sets is presented to help overcoming these difficulties.

2 Background of the Management Simulation Game FuSi

The educational goal of our development activities is to offer students of business computing a computer-based general management simulation game as part of their practical education. We extend previous work from the first author's prior research group at the University of Göttingen [2]. Three components make up the practical game: a) the players that take periodical management decisions based on feedback information from the simulation, b) the game leader who adapts the simulation, starts and supervises the game, discusses the final results with the participants, and c) the simulation model itself that takes the players decisions and converts them to market results (feedback) at the end of each playing period. The feedback to the players comes in the forms of electronic reports. Figure 1 gives an overview of the main components and relationships in the simulation model of our management game FuSi (**F**uzzy **S**imulation). A company is represented by a group

of players who take the position of the company's management. During each playing period they take decisions in the following business areas: procurement (raw material and machines), production (volume, maintenance, environmental investments), sales (price, marketing, customer service), personnel (employment, dismissal at different levels of qualification), finance (credits, cash management).

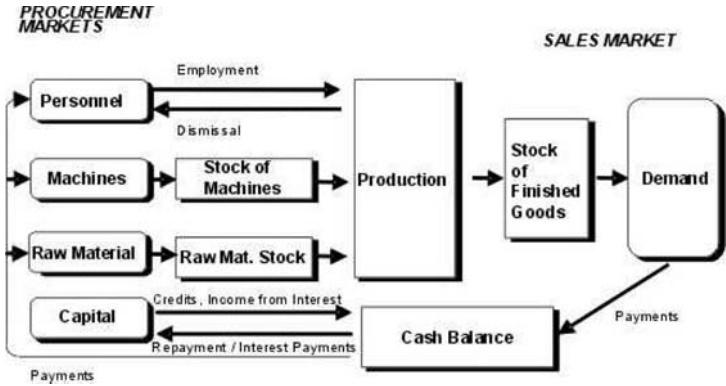


Fig. 1. Overview of simulation model in management game FuSi

Technically, the simulation model consists of two components. One represents the quantitative parts of the model, implemented with mathematical equations in VisualBasicTM under MS AccessTM. The other incorporates the qualitative parts and is realized with the tool Fuzzy Control ManagerTM in the form of a rule-based system. Both parts of the model communicate via the Dynamic Data Exchange (DDE) protocol. Finally, all results of the simulation are integrated in the VisualBasicTM framework for a consolidated output to the participants of the management game.

Many business figures, such as price, sales volume and production costs have a quantitative character and can, therefore, be modeled with traditional equations. Others, like the product image or the credit rating of the company are of a qualitative nature. Also many important relationships are vague and difficult to model with traditional approaches, such as the influence of employee qualification on product quality and scrap volume, or the connection between product image and sales. Fuzzy sets offer an elegant way to model such qualitative information as is outlined in the following section in the context of credit rating.

3 Modeling of Qualitative Aspects: the Example of Credit Rating

Fuzzy rules and inference mechanisms are used in different parts of the simulation model where qualitative information is important. To demonstrate the general procedure, the example of credit rating is described in this section. In this particular sub-application the credit rating of a simulated company is determined at the beginning of each playing period.

Financial and non-financial performance measures are important inputs to the credit rating in our simulation. The rating in turn influences the credit conditions a company faces in the following period of the game. The model of credit rating used here is based on work by Bagus [3] and Rommelfanger [4]. It takes the form of an hierarchical inference system, as outlined in Figure 2. With the product image and the management image two qualitative inputs enter the system. Their values are determined elsewhere in the model through a fuzzy rule-based system that is analogous to the one described here for credit rating. The rest of inputs on the right hand side of Figure 2 are quantitative and mainly taken from the balance sheet and the profit and loss statement of the simulated company at the end of the previous playing period.

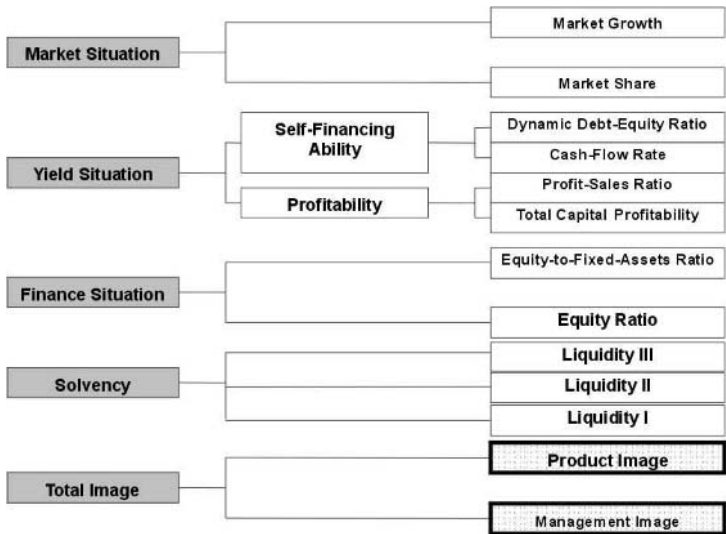


Fig. 2. Components of credit rating in our model

In real life two types of uncertainty occur in the context of credit rating [3, 4]: a) informational uncertainty, meaning that many factors are relevant for the decision, and b) intrinsic uncertainty, meaning that verbal descriptions and expert judgements play an important role. The term "fuzziness" embraces both types of uncertainty [5].

While most of the inputs are quantitative, the relationships between elements are vague and subject to fuzziness. Rommelfanger [4] points out that personal experience, captured in expert rules like the following example, are employed during the process of hierarchical inference in practical credit rating:

IF Profit-Sales Ratio is MEDIUM AND Total Capital Profitability is HIGH, THEN Profitability is MEDIUM.

This suggests a fuzzy rule-based system as an adequate means to model the hierarchical aggregation. In our case, the rules rely on previous field research done by Bagus [3] and Rommelfanger [4] as well as asking credit rating experts in banks. It should be noted that rules obtained in this way may possibly be misleading, unless some form of objective verification is found.

In a first step, the quantitative inputs must be fuzzyfied by mapping them on linguistic terms (such as "low", "medium", "high"). The linguistic terms are modeled as fuzzy sets with normalized degrees of membership in the interval $\mu \in [0, 1]$ ¹. An example is given in Figure 3 for the fuzzyfication of the cash-flow rate with three linguistic terms, modeled as fuzzy sets with trapezoidal membership functions, following Bagus' suggestions [3]. Here, a cash flow of 1.5 % leads to positive degrees of membership both for the linguistic term "low cash-flow" ($\mu = 0.25$) as well as for "medium cash-flow" ($\mu = 0.65$). This suggests that different expert rules are simultaneously activated (to varying degrees) in the process of aggregation to the next-level measure "self-financing ability".

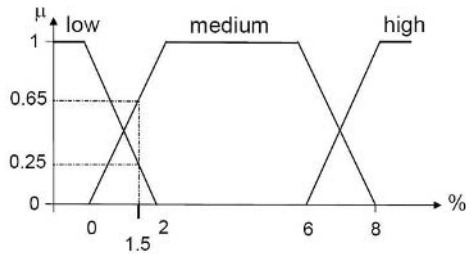


Fig. 3. Fuzzy assessment of the non-fuzzy cash-flow rate 1.5% (example)

Table 1 gives the fuzzy rule-base used in the aggregation of the inputs "cash-flow" and "debt-equity ratio" to the intermediate result "self-financing ability". Results from different active rules are accumulated through the well-established max-min inference operator. The general procedure of inference is similar for all inputs and aggregation levels in Figure 2. In a final step, the results for the five main components on the left hand side of Figure 2 are

¹ For an introduction to the concept of fuzzy sets and fuzzy set theory see [5, 6].

integrated via another rule-base into the final credit rating of the company. The structure of the fuzzy rule-based system that determines the credit rating

Table 1. Rule-base for the assessment of self-financing ability

Cash-Flow Rate (Input)	Debt-Equity Ratio (Input)	Self-Financing Ability (Output)
low	low	low
low	medium	low
low	high	medium
medium	low	low
medium	medium	medium
medium	high	medium
high	low	medium
high	medium	high
high	high	high

based on expert knowledge is depicted in Figure 4. It is equivalent to the hierarchical form of a fuzzy controller as used in technical applications of fuzzy set theory. Based on the fuzzyfication of inputs, the inference mechanism then

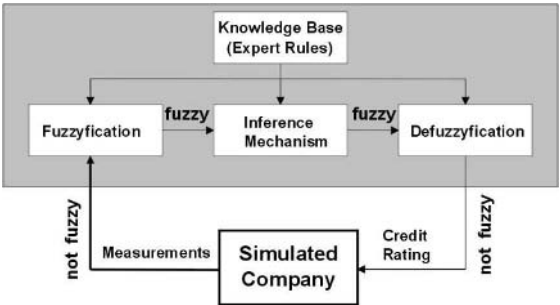


Fig. 4. Structure of the fuzzy rule-based system used for credit rating in FuSi

determines which rules from the knowledge base are activated. The rule base corresponds to a model of the qualitative relationships between the various influencing factors on credit rating. Interestingly, fuzzy rule bases need neither be complete nor free of contradictions to work properly [5]. Furthermore, rules can be weighted to express varying degrees of confidence. Results for the five main components that make up the credit rating (left hand side of Figure 2) are determined through a cascade of rules². The output is converted via center-of-gravity defuzzyfication to a non-fuzzy number that represents the credit rating.

We have used the same principle to model other qualitative aspects in FuSi, such as the concept of "product image" that depends on quantitative decisions in areas like product price, advertising, quality of machines and personnel. The

² For a more detailed account of the working of a fuzzy controller see [5].

qualitative aspects in turn take an influence on the quantitative model part. Image, for instance, influences sales. A good image will make it possible to sell the product at a better price. The credit rating result is used to determine the credit conditions for a company in the next round of the game etc.

4 Conclusion and Current Extensions

Traditional approaches for developing management games suffer from a neglect of qualitative information and vague relationships. Fuzzy rule-based systems offer an elegant way to model such qualitative components. The integration of quantitative and qualitative parts in a combined model then leads to a more realistic representation of the real-world situation. Consequently, modeling qualitative information with fuzzy sets can help to improve the educational value of management games.

Another interesting extension of traditional management simulations is currently under development. Here, we use software agent technology to arrive at an automatic opponent to human players for this game. The agent will be able to learn and improve its performance over time. The basic idea is similar to that of chess computers where a single player finds an automated "intelligent" counterpart. Next to educational goals, an objective of this development is to have a test case for different machine learning strategies in a fairly realistic and complex environment.

References

1. Rohn, W.E. (1995) Einsatzgebiete und Formen des Planspiels. In: Geilhardt, T.; Mühlbradt, T. (eds.): Planspiele im Personal- und Organisationsmanagement, Göttingen: Verlag für angewandte Psychologie
2. Tietze, M.; Nissen, V. (1998) Unscharfe Modellierung im PC-basierten Fuzzy Unternehmensplanspiel FUPS. In: Biethahn, J.; Hönerloh, A.; Kuhl, J.; Leisewitz, M.-C.; Nissen, V.; Tietze, M. (eds.): Betriebswirtschaftliche Anwendungen des Softcomputing, Wiesbaden, Vieweg, pp. 341-354
3. Bagus, T. (1992) Wissensbasierte Bonitätsanalyse für das Firmenkundengeschäft der Kreditinstitute, Frankfurt/M., Peter Lang
4. Rommelfanger, H.J. (1997) Regelbasierte Entscheidungsunterstützung mit Fuzzy-Logik. In: Biethahn, J.; Hönerloh, A.; Kuhl, J.; Nissen, V. (eds.): Fuzzy Set-Theorie in betriebswirtschaftlichen Anwendungen, München, Vahlen, pp. 175-189
5. Zimmermann, H.-J.; Angstenberger, J.; Lieven, K.; Weber, R. (eds.) (1993) Fuzzy Technologien: Prinzipien, Werkzeuge, Potentiale. VDI, Düsseldorf
6. Zimmermann, H.-J. (2001) Fuzzy Set Theory - and Its Applications, 4. Aufl., Boston: Kluwer

Schedule This - A Decision Support System for Movie Shoot Scheduling

Felix Bomsdorf, Ulrich Derigs, and Olaf Jenal

Department of Information Systems and Operations Research (WINFORS),
University of Cologne, Germany
`bomsdorf@uni-koeln.de`
`ulrich.derigs@uni-koeln.de`
`olaf.jenal@uni-koeln.de`

Summary. Creating a movie shoot schedule is an important part of the movie production process. Even for a small movie project already 50 activities requiring 130 resources such as different actors, director, team, special effects and locations etc. have to be scheduled respecting complex constraints which may be imposed on single resources as well as on every activity.

We formulate the movie shoot scheduling problem (MSSP) using and extending the $\alpha|\beta|\gamma$ -scheme suggested by Demeulemeester and Herroelen and we present a prototype of a Decision Support System for MSSP based on a meta-heuristic approach for generating operational schedules.

1 Project Scheduling in the Movie Industry

Project scheduling is an important part of the movie production process. A movie is shot following a shoot schedule, which is based on the movie script. Each scheduled activity is related to a scene of the movie script. It requires a large number of resources of different renewable resource types. There exist only very few hard order-constraints, thus, from a theoretical point there are many feasible schedules with a highly varying quality.

The scheduling process is not only performed at the beginning of the shooting project, but rescheduling is performed, whenever resources become unavailable or other disruptions appear.

State of the Art. So far, the shooting schedule has always been created manually. There are two software systems, which only help to organize some of the data related to the schedule, but not all. One is the system EP Scheduling, formerly known as Movie Magic Scheduling, the other one is only known in Germany and called SESAM Drehplan. The systems never check the schedule during the manual creation for feasibility nor is there any sophisticated functionality to create a schedule automatically or to store the data -besides

the blocking times in SESAM Drehplan- semantically, which would be the first step to automate some processes. Thus, they both lack decision support functionality.

Related Research. We know of only two papers which address the MSSP. One presented at the MIC 2005 ([2]) only proves that a simplified version of the problem obtained by relaxing all the complex domain specific constraints is NP hard and can be solved as a bin packing problem. Fink and Voss ([3]) use a toy example of a MSSP to illustrate the nature of discrete optimization/sequencing problems.

The Challenge of Movie Shoot Scheduling. A system, which implements a model of the MSSP, can highly reduce the complexity of creating a movie shoot schedule. It can facilitate the scheduling tasks at three different levels:

- Automated (re-)scheduling
- Interactive (re-)scheduling
- Manual (re-)scheduling with constraint checking

Our decision support system *Schedule This* implements the movie shooting model and allows better and faster scheduling, thus, the whole process of movie shoot scheduling can be handled more efficiently.

2 The Movie Shoot Scheduling Problem

Basically two steps must be performed in order to create a movie shoot schedule, first a work breakdown structure (WBS) must be created, secondly the items of the WBS must be scheduled, either manually or automatically.

2.1 Creating the WBS

The WBS is based on the movie script. For each scene or shot an item is created following a specific procedure. The standardized structure of a movie script simplifies this procedure: Of each scene or shot the headline in capital letters denotes the location, the atmosphere¹, the duration of this part in the final movie and the number of the scene. During this process the required time for the shooting of the scene will be calculated based on the content of the scene. In the body text of the scene the other information, e.g. the actors - written in capital letters -, props and cars can be found.

¹ E/N = exterior/night, i.e. the scene plays outside during night. Usually each kind of atmosphere is connected to a special color, thus, it can be easily identified by the scheduler. E.g. E/N could be dark blue and I/D (interior/day) could be yellow.

2.2 The Process of Scheduling

The process of scheduling is based on the items² of the WBS, which were created by analyzing the movie script. After creating the WBS, the resulting items have to be scheduled respecting the constraints and objectives which are described in the following paragraphs. The process can be performed in three different modes: manual, automated and interactive (re-)scheduling.

Properties and Constraints of the WBS Items. Each item specifies the duration of the related activity, the required resources (location, actors, props etc.) and the atmosphere, i.e. the preferred execution period (daily time windows corresponding to the atmosphere (day/night, interior/exterior)). Also, soft constraints, which are evaluated in the objective function, are properties of the items. They are considered, if the movie shall be shot in continuity.

Constraints of Resources. The most important constraint of resources is the working time, which is specified by three different characteristics: the maximum available working hours in a row, the daily earliest start/latest end and the minimum idleness. Also blocked times, which might refresh or consume the available working hours of a resource, have to be considered. Thus, no regular shifts or working days can be modelled during the scheduling process, although a calendarized problem must be solved.

Objective Function. One rule of thumb, which underlies the objective function, is 'Do not change locations, whenever it can be avoided.', but also other goals have to be considered during the optimization. Optimizing the resource availability performance (RAP), which consists of renting costs (RAC) and resource availability quality (RAQ), i.e. the perceived quality of a schedule for each resource is another main scheduling objective. Here the RAQ can be approximated by the length and number of gaps as well as by the amount of days involved in the movie shooting for each individual resource. Moreover, two 'creative' nonregular measures of performance exist: to shoot in continuity³ and to shoot in emotional continuity⁴.

Critical Factors. Three types of constraints are the critical factors, which have the largest effect on the feasibility of schedules: the working time especially of child actors, the blocked times of resources and the constraint that each scene must be shot according to the atmosphere, e.g. E/N (exterior night) must be shot during night, where I/N (interior night) can be shot at any time due to the reason that this atmosphere inside can easily be faked.

² I.e. scenes or shots.

³ I.e. if scene A plays in the movie script on July 6th 2005 and scene B plays on September 24th 2005, scene A should be shot before B.

⁴ I.e. shoot in the order of scenes, which is given by the movie script, e.g. a scene on page 7 should be shot before a scene on page 24

2.3 The MSSP as RCPSP

Following the $\alpha|\beta|\gamma$ -scheme suggested by Demeulemeester and Herroelen (see [1]) the MSSP can be classified as a special case of RCPSP. The scheme is closely related to machine scheduling. α denotes the resource environment, β the activity characteristics and γ the objective function. In the next paragraphs the most important domain specific characteristics of the MSSP, which make it differ from the standard RCPSP, will be described.

Resource Environment α . The previously mentioned new resource types with generic working hours and blocked times as well as the fact that resources are almost completely outsourced and rented have a high impact on the schedule and the objective function. The resources are already assigned to activities and cannot be replaced, thus, the schedule depends significantly on the availability of individual resources. This cannot be changed by strategies, as for instance keeping a backup of scarce resources.

Activity Characteristics β . There exist several (daily) time-windows for activities due to their atmosphere, but only a few hard precedence relations. For MSSP a huge amount of feasible schedules exists in contrast to standard RCPSP, where a large number of precedence relations substantially reduces the number of feasible schedules.

Objective Function γ . The most important measure of performance is the number of changes of locations, which highly impacts other measures, e.g. organizational effort, resource availability cost and travels. Two 'creative' non-regular measures of performance exist: to shoot in continuity⁵ and to shoot in emotional continuity⁶.

These characteristics result in complicated resource profiles. For MSSP resource profiles are highly dependent on the schedule, since there are no fixed working shifts etc..

3 'Schedule This' - a DSS for the MSSP

Our system *Schedule This* implements the movie shoot scheduling model. In the following section we will outline the algorithmic basis of the model component and the user interface.

⁵ I.e. if scene A plays in the movie script on July 6th 2005 and scene B plays on September 24th 2005, scene A should be shot before B.

⁶ I.e. shoot in the order of scenes, which is given by the movie script, e.g. a scene on page 7 should be shot before a scene on page 24

3.1 The Model Component - Applying GIST to MSSP

The GIST⁷ Approach is based on the concept of indirect (evolutionary/local) search (see [5]). It has been developed to support rapid and flexible prototyping of DSS solutions in unstructured problem domains (see [4] for an overview). It utilizes an auxiliary search space S^{aux} of simple and general representations. Transformation of solutions in S^{aux} into solutions of the original problem is done by a decoder module $D : S^{aux} \rightarrow S$. The search in S^{aux} is conducted by a general problem independent metaheuristic local search engine. The decoder can be seen as a construction heuristic on S where the encoding $s' \in S^{aux}$ generates/determines the instructions or assignments. To be computationally fast the greedy construction principle is used. It consists of the following modules

- Ordering/sequencing input from the MLS Engine
- Problemspecific decision rules transforming the problem's objective into local strategies applied to encodings
- A constraint checker to control feasibility

To apply the GIST approach to an optimization problem it is necessary to perform the following design steps:

1. Specify a representation of problem solutions in the auxiliary search space S^{aux}
2. Define suitable operators generating neighbor solutions.
3. Implement a greedy decoding procedure transforming a given solution $s \in S^{aux}$ into a solution in the original search space S .

In the following we describe the design of these components in *Schedule This*.

Representation. A solution in S^{aux} is encoded by a permutation $P = a_1, \dots, a_n$ of activities for which feasible start and endtimes need to be found. Additionally the order of the elements in P corresponds to a topological sorting of a precedence graph G which is constructed on the hard order constraints of the MSSP.

Operators. In each iteration we randomly apply one of the following operators by which the property of a topological sorting is respected.

- $Swap(a,b)$: for elements not involved in precedence relations a random swap of two elements (a,b) in P is performed
- $Move(a)$: element a is moved to a new position in P without breaking the topological order constraint in P

⁷ GIST - Indirect Search with Greedy Decoding Technique

Decoder. The decoder is implemented in terms of a deterministic insertion procedure. It schedules activities according to the sequence in the given permutation P . It starts with an empty schedule and inserts activities for which start and end times have been already fixed by the planner in a preprocessing step. Then for each activity in P a number of candidate positions are calculated with respect to resource availability and further temporal constraints e.g. earliest start, latest end and timewindows. Among those the best candidate position according to the decision rule is chosen and all affected resources are allocated. After all activities have been scheduled by the decoder the solution is evaluated and its value is reported to the MLS Engine. In the implementation of *Schedule This* local search is controlled by a threshold accepting algorithm (see [6]).

3.2 The Scheduling Interface

Schedule This facilitates the scheduling process. It enables the scheduler to create the WBS with the constraints of the scenes and the resources (see 2.2). Then the scheduler can use the WBS as a basis for manual scheduling, where the constraints are automatically checked, thus, he can be sure that the schedule is feasible. If he breaks a constraint, the system displays a warning message and gives detailed information about the problem. As an alternative to manual scheduling the system can automatically create a whole schedule respecting all constraints using the method described in 3.1.

During the manual scheduling and re-scheduling process, the system provides interactive decision support. It can propose activities to be scheduled after a certain activity in the current schedule based on measures which optimize the objective function, e.g. minimizing the changes in locations etc.. The system respects all constraints during the interactive scheduling and offers natural language explanations for the proposed allocations instead of the numerical measures. These functions simplify and accelerate the process of manual scheduling and re-scheduling.

4 Experiences and Benefits of Using *Schedule This*

To test the system we used two real world problem instances. The first one was a german TV play. It was 90 minutes long and took 24 shooting days at about 30 different locations. The movie script had 146 pages and described 89 scenes, of six different types of shooting (inside day/outside night/etc.). The team consisted among others of two main actors and about 10 supporting actors. In total we had to schedule about 100 activities using in total 190 different resources.

The schedule for the second instance was already created using our software system *Schedule This*. The production team used our system throughout the whole shooting, thus, they proved that it is suitable to the needs of movie

productions. It was an independent movie, 60 minutes long. The shooting took place during 17 shooting days in Berlin and the surrounding of Berlin at about 20 different locations. The movie script described 38 scenes of six different types. The team consisted among others of three main actors and two supporting actors. In total about 50 activities were scheduled using 130 different resources.

For both instances a feasible solution was computed within seconds, and a reasonable good schedule was computed within five to ten minutes. These good schedules only required small changes by the planner and, thus, saved hours of work for creating an initial schedule and improving this schedule manually.

Schedule This saves many working hours due to automated and interactive (re-)scheduling with automated constraint checking. Thus, compared to manual scheduling schedules of higher quality can be created in shorter time.

References

1. Demeulemeester, Erik L. ; Herroelen, Willy S. (2002) Project Scheduling: A Research Handbook. Kluwer Academic Publishers, Boston Dordrecht London
2. Jensen M. B., Kaufmann M., and Zachariasen M. (1999): "Movie Shoot Scheduling". In: *MIC 2005*, 551–556.
3. A. Fink und S. Voß: HOTFRAME - Heuristische Lösung diskreter Planungsprobleme mittels wiederverwendbarer Software-Komponenten. *OR News* 4 (1998), 18 - 24.
4. Derigs, U., Döhmer, T., Jenal, O. (2005), Indirect Search With Greedy Decoding Technique (GIST) - An Approach for Solving Rich Combinatorial Optimization Problems. *MIC 2005*.
5. Gottlieb, J. (2001) Evolutionary Algorithms for Constrained Optimization Problems. Shaker Verlag, Aachen
6. Dueck G., Scheuer, T. (1990) Threshold Accepting: A General Purpose Optimization Algorithm Appearing Superior to Simulated Annealing. *Journal of Computational Physics* 90:161–175.

A Framework for Truth Maintenance in Multi-Agent Systems

Brett Bojduj¹, Ben Weber¹, and Dennis Taylor²

¹ California Polytechnic State University, San Luis Obispo, U.S.A.

bbojduj@calpoly.edu

bgweber@calpoly.edu

² CDM Technologies, Inc., San Luis Obispo, U.S.A.

djtaylor@cdmtech.com

Summary. Maintaining logical consistency in a knowledge base that receives asynchronous events from collaborative agents poses a challenge, due to multiple agent perspectives of the knowledge base. Facades and filters further complicate this problem by distorting the definition of consistent knowledge. Therefore, maintaining logical consistency of the knowledge base requires an infrastructure for handling truth maintenance. This paper presents a generic, object-oriented framework for truth maintenance in collaborative multi-agent systems. The core of the framework is an agent that autonomously reasons on system events, thus guaranteeing the integrity of the knowledge base independent of external agents. Specialization to a particular domain is achieved through the description of tests that verify the consistency of the knowledge base. This paper shows an example of this approach in a real-world, multi-agent system and discusses performance and maintainability in such a system.

1 Introduction

A challenge for large software systems is maintaining logical consistency between collaborative agents operating asynchronously to solve a problem. Multi-agent systems may use a divide-and-conquer approach to solve a problem [4]. They can do this by assigning different aspects of the problem to different agents. Each agent in the system may have a different perspective of the problem and the definition of consistency may vary between agents. Therefore, multi-agent systems require that solutions are consistent across all agent perspectives. This paper presents an external truth maintenance agent as a technique for maintaining consistency between collaborative agents.

Multi-agent systems have been used to solve constraint satisfaction problems [7]. For large, highly-constrained problems, implementation of an agent that simultaneously solves all constraints becomes impractical. A common solution is to construct collaborative agents that subdivide the problem into

smaller problems and share solutions. Decomposing the problem requires that solutions shared between two agents are consistent with both agents' perspectives of the problem. Consistency can be achieved if agents communicate through a shared knowledge base that is monitored by an external truth maintenance agent. The truth maintenance agent verifies that solutions in the knowledge base are consistent with respect to all constraints in the problem. This alleviates individual agents from the burden of maintaining consistency for the entire solution.

This paper explores the use of a truth maintenance agent to maintain logical consistency of a knowledge base in a global logistics and scheduling application. The task of planning is distributed between several collaborative agents and the consistency of the knowledge base is verified by an external truth maintenance agent. The presented technique enables metaheuristic-based agents to generate consistent solutions in a near real-time planning environment.

2 Truth Maintenance Systems

Most agents require logically consistent information in order to perform their reasoning functions. In dynamic environments, additional information is added to the problem definition while the system is running, which may conflict with current solutions. If logically inconsistent data enters a system, existing solutions must be marked as invalid or revised in order to allow agents to reason accurately.

Doyle proposed a truth maintenance system to handle cases when logically inconsistent data enters a system [2]. Each set of data in the system is annotated with a node that specifies the consistency of the data. When new information enters the system, every node is updated to reflect the new problem definition. Huhns and Bridgeland extended the truth maintenance system to enable consistency in multi-agent systems [5]. When new information enters the system, every agent verifies that its shared data is consistent. This validation process becomes a bottleneck, but is required for blackboard architectures. If the knowledge base becomes compromised, then agents in the system communicating through the blackboard will be unable to properly perform their reasoning tasks.

Traditional truth maintenance systems guarantee consistency of a knowledge base, but are unsuitable for real-time systems. Elkan proved that truth maintenance systems that deal with logically inconsistent information are NP-complete [3]. Therefore, systems must compromise global consistency in order to operate in real-time environments. The amount of computation required to enforce consistency can be reduced if agents consider different aspects of the problem. Bertel et al. presented the concept of "aspect filters" which can be applied to truth maintenance systems [1]. When new data enters the system, consistency is maintained within only the affected aspects of the knowledge

base. If an aspectualization approach is used, then agents working on different aspects of a problem are not required to be logically consistent with respect to each other. This technique lends itself well to metaheuristic-based agents that perform reasoning with a shared knowledge base.

3 Definition of Consistency

The purpose of a truth maintenance system is to verify the consistency of a knowledge base. We define the consistency of a knowledge base, C_{KB} , to be:

$$C_{KB} = F_V(O, C) \quad (1)$$

where O is the set of objects in the problem definition and C is the set of defined constraints. The verification function, F_V , checks that all constraints are satisfied by the set of objects. The knowledge base is defined as consistent when all constraints in the problem definition are satisfied. In our implementation, we apply this definition of consistency to a real-time vehicle routing and scheduling problem.

4 Truth Maintenance Agent

This paper presents a technique for maintaining consistency through the use of an agent that runs independent of the rest of the system. The purpose of the truth maintenance agent is to verify that solutions shared in a knowledge base are globally consistent. Additionally, the agent checks the consistency of the knowledge base when new data enters the system. The truth maintenance agent alleviates other agents from checking the validity of objects in the knowledge base. This design facilitates communication between agents operating in a distributed environment by enabling agents with different perspectives to communicate consistent solutions.

When a solution is posted to the knowledge base, the truth maintenance agent checks the validity of the solution. During this process, other agents in the system may utilize the solution. If the solution is determined to be consistent, then the truth maintenance agent performs no further actions. However, if the solution is logically inconsistent, then the truth maintenance agent marks the solution as invalid and broadcasts an alert to the other agents in the system. Agents operating on invalid solutions stop their reasoning processes and check for revised solutions from the knowledge base or generate new solutions.

The truth maintenance agent verifies the consistency of solutions using an aspect-oriented approach. The verification function is composed of a series of tests, which check various aspects of objects in the knowledge base. The agent contains a mapping of external events to tests that specifies which tests

to run when new data enters the system. This approach allows the truth maintenance agent to maintain consistency without exhaustively testing the knowledge base. However, an exhaustive approach is used when new solutions are posted to the knowledge base, because the truth maintenance agent must verify that new solutions are globally consistent.

The truth maintenance agent provides a technique for maintaining consistency in a system, but there are several tradeoffs. The agent runs independently of the system, which may result in a delay between the time when a solution is posted and the time when the validity of the solution is verified. Also, the truth maintenance agent does not monitor the internal states of other agents in the system. Therefore, the main disadvantage is that agents in the system may temporarily operate on invalid data. However, several benefits are gained through the use of a truth maintenance agent. Agents in the system are not required to translate between different perspectives, reducing code complexity. The truth maintenance agent can also be applied to dynamic problem domains. Agents continue solving a problem when a change occurs and are informed by the truth maintenance agent if the knowledge base becomes inconsistent.

5 Implementation in TRANSWAY

The truth maintenance agent was developed to facilitate communication between collaborative agents in a logistics and scheduling application called TRANSWAY, developed by CDM Technologies, Inc. The purpose of TRANSWAY is to provide execution plans for military distribution of supplies at strategic and operational levels of planning. The problem is represented as a constrained vehicle routing problem with multiple types of conveyances and transshipments. The main challenge for the TRANSWAY system is providing up-to-date plans, because new data frequently enters the system.

The task of planning is distributed between three agents in TRANSWAY: the routing agent, the planning agent, and the impediment agent. The routing agent maintains the shortest paths between all locations in a scenario and enforces conveyance-range constraints. The planning agent uses a tabu search algorithm to generate solutions using paths from the routing agent and handles scheduling constraints. The impediment agent verifies that solutions discovered by the planning agent do not require conveyances to travel through impediments, such as sandstorms. A multi-agent approach enables decomposition of the problem, but prevents individual agents from determining if solutions are globally consistent.

The most common usage of the truth maintenance agent is maintaining consistency when new information about impediments enters the system. If a weather impediment is created or modified, the truth maintenance agent performs the tests mapped to these events to determine the validity of the knowledge base. There are two ways an impediment can invalidate a solution: a

light impediment limits the cruising speed of vehicles and results in scheduling conflicts in the solution, while a severe impediment causes routes to become impassible.

When a weather impediment is created or modified, the truth maintenance agent checks that conveyances in the solution are not scheduled to traverse through impeded routes. If a conveyance is scheduled to pass through an impeded route that makes its scheduled delivery impossible, the truth maintenance agent marks the delivery as invalid and sends an alert to the routing and planning agents. The routing and planning agents are responsible for generating a new solution that is consistent with the updated scenario definition. The routing agent validates the solution by generating routes that traverse around the impediment. Afterwards, the planning agent removes the invalid delivery and reschedules the delivery using updated routes from the routing agent.

The truth maintenance agent has enabled the TRANSWAY system to solve logistics problems with several constraints. Computationally expensive processes required to verify the consistency of solutions are delegated to the truth maintenance agent, allowing other agents in the system to continue solving a scenario. Implementation of the truth maintenance agent in TRANSWAY has demonstrated the feasibility of maintaining consistency between collaborative agents in a near real-time planning environment.

6 Conclusion and Future Work

Maintaining consistency of a knowledge base in a multi-agent system is computationally expensive. This process becomes impractical when agents must interact in a real-time environment. Therefore, it is necessary to tradeoff global consistency for improved performance. We have presented a technique for maintaining consistency of a knowledge base using an external truth maintenance agent. The truth maintenance agent facilitates communication between agents with different perspectives and enables agents to operate in near real-time environments.

The main drawback of the presented truth maintenance technique is that agents may operate on invalid information for large periods of time. Therefore, agents should not be completely reliant on the truth maintenance agent for validation of the knowledge base. Taylor et al. presented a technique for generating consistent solutions to dynamic problems for a single metaheuristic-based agent [6]. Future work includes leveraging this technique with the truth maintenance agent to achieve improved agent performance in highly dynamic environments.

References

1. Bertel S, Freksa C, Vrachliotis G (2004) Aspectualize and conquer in architectural design. In: Gero JS, Tversky B, Knight T (eds) *Visual and Spatial Reasoning in Design III*. Key Centre of Design Computing and Cognition; University of Sydney 255–279
2. Doyle J (1979) A truth maintenance system. *Artificial Intelligence* 12(3):231–272
3. Elkan C (1990) A rational reconstruction of nonmonotonic truth maintenance systems. *Artificial Intelligence* 43(2):219–234
4. Ephrati E, Rosenschein JS (1994) Divide and conquer in multi-agent planning. *AAAI '94: Proceedings of the twelfth national conference on artificial intelligence* 1:375–380
5. Huhns MN, Bridgeland DM (1991) Multiagent truth maintenance. *IEEE Transactions on Systems, Man, and Cybernetics* 21(6):1437–1445
6. Taylor D, Weber B, Bojduj B (2006) A tabu search framework for dynamic combinatorial optimization problems. *Proceedings, Integrated Design and Process Technology, IDPT-2006* 663–668
7. Yokoo M (2001) *Distributed constraint satisfaction: foundations of cooperation in multi-agent systems*. Springer-Verlag, London

Multi Criteria Decision Theory

Preference Sensitivity Analyses for Multi-Attribute Decision Support

Valentin Bertsch, Jutta Geldermann, and Otto Rentz

Institute for Industrial Production (IIP), University of Karlsruhe (TH), Germany
valentin.bertsch@wiwi.uni-karlsruhe.de

Summary. Contributing to transparency and traceability of decision making processes and taking into account the preferences of the decision makers, multi-criteria decision analysis (MCDA) is suitable to bring together knowledge from different disciplines and fields of expertise. The modelling of the decision makers' preferences is a crucial part of any multi-criteria analysis. In multi-attribute value theory (MAVT), preferential information is modelled by weighting factors (i.e. *inter-criteria* comparisons) and value functions (i.e. *intra-criteria* preferences). However, the uncertainties associated with the determination of these preferential parameters are often underestimated. Thus, the focus of this paper is the description of an approach to explore the impact of simultaneous variations of these subjective parameters. Special attention is paid to the consideration of variations of the value functions' shapes. The aim of the presented methods is to facilitate the process of preference modelling and to comprehensibly visualise and communicate the impact of the preferential uncertainties on the results of the decision analysis.

1 Introduction

A decision making process in practice is subject to various sources of uncertainty. The occurring uncertainties can be classified in many different ways, see for instance [8, 3, 5]. According to their respective source, a distinction can be made between *data uncertainties* (uncertainties of the input data to a model), *parameter uncertainties* (uncertainties related to the model parameters) and *model uncertainties* (uncertainties resulting from the fact that models are ultimately only simplifications/approximations of reality). *Model uncertainties* are difficult to quantify and can also be regarded as inherent to the nature of any model. They are thus not considered in this paper. An approach to model, propagate and visualise *data uncertainties* within a decision support system has for instance been proposed in [5].

In this paper, special attention is paid to the uncertainties associated with the *preference parameters* of a MCDA model. These uncertainties are usually examined by means of “common” sensitivity analyses where the best-known

approach allows to examine the effects of varying a weighting parameter of a MCDA model. A major drawback of this approach is that it is limited to varying one (weighting) parameter at a time. The use of Monte Carlo simulation for the consideration of simultaneous variations of the weights of a decision model is described in [1]. This paper presents an extension of the approach to examine the impact of simultaneous variations of the value functions' shapes and their domains' boundaries. In addition, a combined consideration of the different types of preferential uncertainties and their respective impacts on the results of the analysis is proposed.

2 Value Function Sensitivity Analysis

In MAVT, the overall performance score of an alternative a is usually evaluated as

$$v(a) = \sum_{j=1}^n w_j \cdot v_j(s_j(a)) \quad (1)$$

where $1 \leq j \leq n$ and n denotes the number of considered attributes, $s_j(a)$ denotes the score of alternative a with respect to attribute j , $w = (w_1, \dots, w_n)$ ($\sum_{j=1}^n w_j = 1$, $w_j \geq 0$ for all j) is the weighting vector and $v_j(x)$ (with $x = s_j(a)$) is the value function to evaluate the performance of alternative a with respect to attribute j . For the latter, the common one-parameter representation for exponential value functions will be used in the following (see also Fig. 1):

$$v_j(x) = \begin{cases} \frac{1 - e^{-\frac{\Delta_j x}{\rho_j}}}{1 - e^{-\frac{x_{max}^j - x_{min}^j}{\rho_j}}}, & \rho_j \neq \pm\infty \\ \frac{\Delta_j x}{x_{max}^j - x_{min}^j}, & otherwise \end{cases} \quad (2)$$

with

$$\Delta_j x = \begin{cases} x - x_{min}^j & \text{for increasing preferences,} \\ x_{max}^j - x & \text{for decreasing preferences.} \end{cases} \quad (3)$$

2.1 Varying the Value Functions' Shapes

In the following, the impact of varying the parameters ρ_j (i.e. varying the shape(s) of the value function(s)) shall be analysed. For this, let $\rho = (\rho_1, \dots, \rho_n)$ denote the vector of the value functions' shape parameters. In order to explore the effect of simultaneously varying the value functions' shapes, we propose to assign an interval $I_j(\rho)$ to each attribute j by determining a lower (ρ_j^l) and an upper (ρ_j^u) bound instead of assigning a discrete value ρ_j :

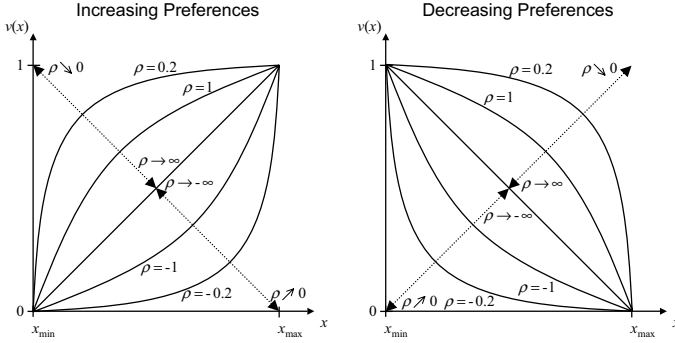


Fig. 1. Value function shapes for different values of the parameter ρ

$$I_j(\rho) = [\rho_j^l, \rho_j^u] . \quad (4)$$

Then, the uncertainties of the intra-criteria preferences can be described by the n -dimensional interval

$$C_\rho = \prod_{j=1}^n [\rho_j^l, \rho_j^u] \times \dots \times [\rho_n^l, \rho_n^u] . \quad (5)$$

In comparison to exploring the impact of simultaneously varying the weights within certain intervals (see e.g. [1]), the procedure is much simpler for value functions since, for the ρ_j , there is no constraint such as that their sum needs to be equal to 1 as for the weights. Monte Carlo Simulation can be used to draw samples of ρ within the intervals. Because of the exponential function in the definition of the value function (cf. equation (2) on page 412), it is not advisable to simply presume a uniform distribution of the elements of the set C_ρ . For instance, there is not much difference in the curvature of the value functions corresponding to the parameters $\rho_j = 5$ and $\rho_j = 10$, whereas the difference is much larger for the functions corresponding to $\rho_j = 0.2$ and $\rho_j = 0.3$. Thus, a logarithmic transformation is performed at first and then, it is assumed that the elements of the transformed set can be considered to be equiprobable. Once the samples are drawn, the approach is in principal straightforward and furthermore computationally easily feasible.

2.2 Varying the Value Functions' Domains

Besides varying the value functions' shapes it is also interesting to investigate the effect of varying their domains' boundaries. In practice, the boundaries are often defined by the minimum and maximum scores actually achieved by the different alternatives (with respect to the considered attributes). By following this approach, theoretically possible better or worse outcomes are neglected.

However, the estimation of reasonable values for these theoretically possible boundaries is a difficult task. Thus, in order to support a decision making team in coping with this task, it is suggested to analyse whether or not the variation of the boundaries has an impact on the results. The procedure is very similar to the one described in section 2.1. Instead of varying the parameters ρ_j , the boundaries x_{min}^j and/or x_{max}^j are varied within certain intervals.

3 A Hypothetical Case Study

The approach will be demonstrated by means of a hypothetical case study from nuclear emergency and remediation management. This case study was originally used for discussion within a moderated decision making workshop in Germany. Details of the case study are described in [6]. The data for this hypothetical case study was generated by RODOS (Real-time Online Decision Support System for Nuclear Emergency Management, see: <http://www.rodos.fzk.de>). In the event of a nuclear or radiological emergency in Europe, RODOS provides consistent and comprehensive decision support at all levels ranging from largely descriptive reports, such as maps of the contamination patterns and dose distributions, to a detailed evaluation of the various countermeasure or remediation strategies and their ranking according to the societal preferences as perceived by the decision makers [4, 2]. For the latter, the multi-criteria decision support tool Web-HIPRE (see [9]) has been integrated into RODOS [6].

The aim of the case study within the moderated workshop was the choice of the „best“ remediation strategy for a contaminated inhabited area after a fictitious release of radioactive material into the atmosphere (from a nuclear power plant). The problem structuring process resulted in an attribute tree showing the overall goal („best“ remediation strategy) as the top criterion which was subdivided into the criteria „radiological dose“ „logistics“ and „costs“ where „radiological dose“ and „logistics“ were split up again resulting in a total of seven attributes. The seven considered countermeasure and remediation strategies ranged from „No action“ (Strategy S1) to „Evacuation, removal of bushes and shrubs, lawn mowing and removal of street surfaces“ (Strategy S7). The strategies S2-S6 comprise actions on a scale in between these two options.

3.1 Value Function Modelling

In order to support a decision maker or a group of decision makers in determining the shape(s) of the value function(s), they are allowed to assign intervals for the parameters ρ_j (which define the shapes) instead of precise values. To demonstrate the described method, the parameters ρ_j for the seven attributes are all varied between 0.2 and 10 (or -10 and -0.2 respectively).

3.2 Selected Results

A prototype was implemented to demonstrate the described approach. Results for 1000 samples of the random vector ρ are shown in Fig. 2. Diagram (a) shows the results for “only” varying the parameters ρ_j . Diagram (b) shows the results when the upper boundaries of the value functions’ domains are additionally varied in an interval between the maximum of the occurring scores and a value augmented by 20%. Diagram (c) shows the effects of varying the weights in addition (see [1] for more details in this regard). Fig. 2 shows that the exclusive variation of the value functions’ shapes does not have a large impact on the results for the considered case study. An additional variation of the boundaries of their domains increases the impact but no rank reversals are observable. For the combined variation of all considered preference parameters, the impact is noticeable and a rank reversal is observable. In this case, strategy S7 has the highest overall performance score in 54% of the drawn samples and strategy S6 gets rated highest in 46% of the cases. Finally, it should be mentioned that strategy S7 gets rated highest in 100% of the cases when “only” the weights are varied.

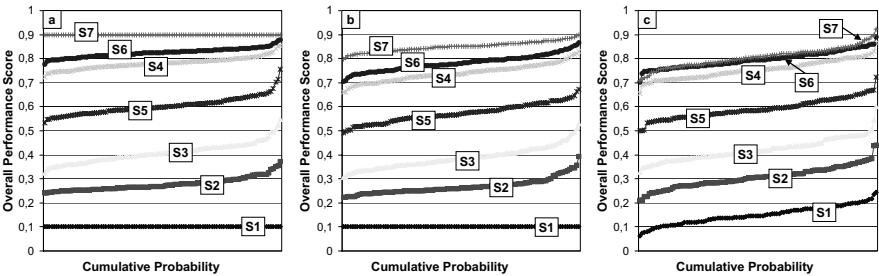


Fig. 2. Selected Results (overall performance scores plotted versus cumulative probability): (a) for “only” varying the parameters ρ_j (left), (b) for additionally varying the upper boundaries (middle), (c) for also varying the weights (right)

4 Conclusions

An approach has been introduced to consider the impact of simultaneous variations of several (preferential) parameters on the results of the decision analysis, aiming at facilitating the process of preference elicitation, exploring the robustness of the analysis and allowing an informative overview and deeper insight into the decision situation. It should be emphasised that the approach is aimed at communicating and visualising the uncertainties associated with the results of the decision analysis by explicitly illustrating their spread - i.e. the ranges in which the results can vary. It would now be interesting, to test

and evaluate the approach in further decision making workshops in order to ensure that it meets the needs of decision making processes in practice.

Possible enhancements of the approach include, for instance, the application of Principal Component Analysis (PCA), i.e. a projection of the (higher dimensional) results on a (two-dimensional) plane with minimal loss of information aiming at providing a rapid overview of the results of decision analyses (see e.g. [10, 7]). Furthermore, decisions are usually not taken at one single point as it is often assumed for reasons of simplicity. In reality, decisions are rather nested in a series of decisions which are related to each other. Thus, the aim is the development of operationally applicable extensions of the existing methods which reflect the sequential and iterative process of decision making in real life.

References

1. V. Bertsch, J. Geldermann, and O. Rentz. Multidimensional Monte Carlo Sensitivity Analysis in Multi-Criteria Decision Support. In F. Mayer and J. Stahre, editors, *Proceedings of the 9th IFAC Symposium on Automated Systems Based on Human Skill And Knowledge, Nancy, France*, 2006.
2. J. Ehrhardt and A. Weiss. RODOS: Decision Support for Off-Site Nuclear Emergency Management in Europe. EUR 19144 EN. *Luxembourg, European Community*, 2000.
3. S. French. Uncertainty and imprecision: Modelling and Analysis. *Journal of Operational Research Society*, 46:70–79, 1995.
4. S. French, J. Bartzis, J. Ehrhardt, J. Lochar, M. Morrey, N. Papamichail, K. Sinkko, and A. Sohler. RODOS: Decision support for nuclear emergencies. In S. H. Zanakakis, G. Doukidis, and G. Zopounidis, editors, *Recent Developments and Applications in Decision Making*, pages 379–394. Kluwer Academic Publishers, 2000.
5. J. Geldermann, V. Bertsch, and O. Rentz. Multi-Criteria Decision Support and Uncertainty Handling, Propagation and Visualisation for Emergency and Remediation Management. In H.-D. Haasis, H. Kopfer, and J. Schönberger, editors, *Operations Research Proceedings 2005*, pages 755–760. Springer, 2006.
6. J. Geldermann, M. Treitz, V. Bertsch, and O. Rentz. Moderated Decision Support and Countermeasure Planning for off-site Emergency Management. In R. Loulou, J.-P. Waaub, and G. Zaccour, editors, *Energy and Environment: Modeling and Analysis*, pages 63–80. Springer, 2005.
7. J. Hodgkin, V. Belton, and A. Koulouri. Supporting the Intelligent MCDA user: A Case Study in Multi-person Multicriteria Decision Support. *European Journal of Operational Research*, 160(1):172 – 189, 2005.
8. M. G. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York, 1990.
9. J. Mustajoki and R. P. Hämäläinen. Web-HIPRE: Global Decision Support by Value Tree and AHP Analysis. *INFOR*, 38(3):208–220, 2000.
10. N. H. Timm. *Applied multivariate analysis*. Springer, New York, 2002.

MCDA in Analyzing the Recycling Strategies in Malaysia

Santha Chenayah¹, Agamuthu Periathamby², and Eiji Takeda³

Faculty of Economics and Administration, University of Malaya, Kuala Lumpur, Malaysia

santha@um.edu.my

Institute of Biological Science, University of Malaya, Kuala Lumpur, Malaysia

agamuthu@um.edu.my

Graduate School of Economics, Osaka University, Japan

takeda@econ.osaka-u.ac.jp

1 Introduction

The outranking analysis has been frequently used to deal with the complex decisions involving qualitative criteria and imprecise data (see [5]).

So far, various variants of and PROMETHEE (Preference Ranking Organization METHod for Enriching Evaluations) have also been widely used for the outranking analysis (see, [1]).

This paper demonstrates a new exploitation procedure based on the eigenvector using the “weighted” in- and out- preference flows of each alternative in the outranking analysis of the recycling strategies in Subang Jaya, Malaysia.

The rest of this paper is organized as follows: In the next section, we shall briefly review the preference flows in a PROMETHEE context and a new exploitation procedure based on the eigenvector using the “weighted” in- and out- preference flows of each alternative in the outranking analysis. Section 3 addresses a case study where we shall introduce the 14 recycling strategies (consisting of different combinations of awareness creation rate and increase in recycling facilities) to raise the current recycling (collecting) rate to achieve the national recycling target of 22% by 2020 in Malaysia. Using several criteria, we conducted an outranking analysis, because qualitative criteria and imprecise data are involved in evaluating the recycling strategies. Concluding remarks are given in the final section.

2 Valued Binary Relation and Preference Flows in PROMETHEE

We begin with a brief review of the partial preorder which is a fundamental concept for ranking(see [3]). Let us consider the set A of n alternatives: $A = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}$.

Let a valued preference relation π on the set A be a function $\pi : A \times A \rightarrow [0,1]$. In what follows, for any ordered pair $(\mathbf{a}_i, \mathbf{a}_j)$, we shall denote by $\pi(\mathbf{a}_i, \mathbf{a}_j)$ a degree to which the relation between $\mathbf{a}_i$ and $\mathbf{a}_j$ holds.

Definition (partial preorder) A valued binary relation π on A is called a partial preorder if it is reflexive and transitive, where π is called reflexive if $\pi(\mathbf{a}_i, \mathbf{a}_i) = 1$ for $i = 1, 2, \dots, n$, and is called transitive if $\pi(\mathbf{a}_i, \mathbf{a}_j) \geq \min\{\pi(\mathbf{a}_i, \mathbf{a}_k), \pi(\mathbf{a}_k, \mathbf{a}_j)\}$ for any $i, j, k = 1, 2, \dots, n$.

Let $g_1, g_2, \dots, g_m$ be m -criteria. Thus, each alternatives $\mathbf{a}_i$ is characterized by a multiattribute outcome denoted by a vector $(g_1(\mathbf{a}_i), g_2(\mathbf{a}_i), \dots, g_m(\mathbf{a}_i))$. In what follows, with no loss of generality, we assume that the decision maker prefers larger to smaller values for each criterion. The valued outranking relation $\pi(\mathbf{a}_i, \mathbf{a}_j)$ of $\mathbf{a}_i$ over $\mathbf{a}_j$ which is a valued preference relation is constructed from the notions of quasi-criterion and pseudo-criterion. (see [1]). Thus, $\pi(\mathbf{a}_i, \mathbf{a}_j)$ represents the intensity of the preference of $\mathbf{a}_i$ over $\mathbf{a}_j$ for all the criteria: the closer to 1, the greater the preference. From a valued outranking relation, a valued outranking graph with nodes corresponding to alternatives and arcs $(\mathbf{a}_i, \mathbf{a}_j)$ having values $\pi(\mathbf{a}_i, \mathbf{a}_j)$ is depicted.

Then, in the PROMETHEE method, the preference out-flow and preference in-flow of each node $\mathbf{a}_i$ are respectively defined by

$$\phi^+(\mathbf{a}_i) = \sum_j \pi(\mathbf{a}_i, \mathbf{a}_j), \quad (1)$$

$$\phi^-(\mathbf{a}_i) = \sum_j \pi(\mathbf{a}_j, \mathbf{a}_i). \quad (2)$$

The higher the preference out-flow and the lower the preference in-flow, the better the alternative.

We now extend the preference flows to the “weighted” preference flows which yield the eigenvalue problem:

$$\lambda \psi^+(\mathbf{a}_i) = \sum_{j=1}^n \pi(\mathbf{a}_i, \mathbf{a}_j) \psi^+(\mathbf{a}_j), \quad i = 1, 2, \dots, n, \quad (3)$$

where λ is a constant and $\psi^+(\mathbf{a}_j)$ is the weight representing the strength of preference of $\mathbf{a}_j$. This implies that it should be better to outrank a „strong,, alternative than a „weak“ one.

Similarly, we define the weighted preference (in-) flows:

$$\lambda \psi^-(\mathbf{a}_i) = \sum_{j=1}^n \pi(\mathbf{a}_j, \mathbf{a}_i) \psi^-(\mathbf{a}_j), \quad i = 1, 2, \dots, n, \quad (4)$$

which implies that it is less serious to be outranked by a „strong“ alternative than by „weak“one. The higher $\psi^+(\mathbf{a}_i)$ and the lower $\psi^-(\mathbf{a}_i)$, the better the alternative. Note that $\psi^+(\mathbf{a}_i)$ and $\psi^-(\mathbf{a}_i)$, in the right hand side of eqs. (3) and (4) are not known in advance but are determined by the eigenvector method.

In what follows, to assure the positive eigenvalue and the corresponding eigenvector whose elements are positive, we replace $\pi(\mathbf{a}_i, \mathbf{a}_j) = 0$ by $\pi(\mathbf{a}_i, \mathbf{a}_j) = \varepsilon$ for any $\mathbf{a}_i, \mathbf{a}_j$, so as to make the matrix Π corresponding to any valued preference relation π a positive matrix, where ε is a sufficiently small positive number so as not to affect the order.

Theorem *Let a valued preference relation π be a partial preorder such that $\pi(\mathbf{a}_i, \mathbf{a}_j) = 1 - \pi(\mathbf{a}_j, \mathbf{a}_i)$ for $i \neq j; i, j = 1, 2, \dots, n$. Then we have,*

$$\begin{aligned} \pi(\mathbf{a}_i, \mathbf{a}_j) &> \pi(\mathbf{a}_j, \mathbf{a}_i) \text{ if and only if } \psi^+(\mathbf{a}_i) > \psi^+(\mathbf{a}_j), & (\psi^-(\mathbf{a}_i) < \psi^-(\mathbf{a}_j)), \\ \pi(\mathbf{a}_i, \mathbf{a}_j) &= \pi(\mathbf{a}_j, \mathbf{a}_i) \text{ if and only if } \psi^+(\mathbf{a}_i) = \psi^+(\mathbf{a}_j), & (\psi^-(\mathbf{a}_i) = \psi^-(\mathbf{a}_j)). \end{aligned}$$

Example Consider the following matrix $\Pi = (\pi(\mathbf{a}_i, \mathbf{a}_j))$ in which π is a partial preorder such that $\pi(\mathbf{a}_i, \mathbf{a}_j) = 1 - \pi(\mathbf{a}_j, \mathbf{a}_i)$ for $i \neq j; i, j = 1, 2, \dots, n$.

$$\begin{array}{c} \mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3 \quad \mathbf{a}_4 \\ \mathbf{a}_1 \begin{pmatrix} 1 & 0.3 & 0.5 & 0.3 \\ \mathbf{a}_2 \begin{pmatrix} 0.7 & 1 & 0.7 & 0.6 \\ \mathbf{a}_3 \begin{pmatrix} 0.5 & 0.3 & 1 & 0.3 \\ \mathbf{a}_4 \begin{pmatrix} 0.7 & 0.4 & 0.7 & 1 \end{pmatrix} \end{pmatrix} \end{pmatrix} \end{pmatrix} \end{array}$$

We get $(\lambda_{max} = 2.4139)$

$$\psi^+ = (\psi^+(\mathbf{a}_i)) = (0.198 \ 0.318 \ 0.198 \ 0.286), \quad \psi^- = (\psi^-(\mathbf{a}_i)) = (0.730 \ 0.452 \ 0.730 \ 0.502).$$

We can see that

$$\begin{aligned} \pi(\mathbf{a}_i, \mathbf{a}_j) &> \pi(\mathbf{a}_j, \mathbf{a}_i) \text{ if and only if } \psi^+(\mathbf{a}_i) > \psi^+(\mathbf{a}_j), & (\psi^-(\mathbf{a}_i) < \psi^-(\mathbf{a}_j)), \\ \pi(\mathbf{a}_i, \mathbf{a}_j) &= \pi(\mathbf{a}_j, \mathbf{a}_i) \text{ if and only if } \psi^+(\mathbf{a}_i) = \psi^+(\mathbf{a}_j), & (\psi^-(\mathbf{a}_i) = \psi^-(\mathbf{a}_j)). \end{aligned}$$

Then the ranking is: $a_2 \succ a_4 \succ a_1 \sim a_3$.

3 MCDA in Evaluating Recycling Strategies in Subang Jaya

3.1 Recycling Strategies

We used the Logistic Curve to estimate the waste growth of Subang Jaya. Furthermore, in our model, substitutability of facilities are treated by fuzzy integral when the facilities increase in number. To achieve the national recycling objective, both increase in recycling facilities and awareness creation is important (see [2]).

Focusing on the two main streams mentioned above, we formulated the following fourteen strategies:

- Strategy 1: {Effect of awareness creation: status quo; recycling facilities: status quo}
- Strategies 2 to 7: {Effect of awareness creation: relatively high ($\lambda = 1\%, 10\%, 15\%, 20\%, 25\%$ and 30%); Recycling facilities: status quo}
- Strategy 8: {Effect of awareness creation: status quo; Increase the number of recycling facilities in 2008: Recycling centre: 0% , Recycling bin: 100% , Recycling Van: 600% , Silverbox: 300% }
- Strategy 9: {Effect of awareness creation: status quo; Increase recycling facilities in 2-phase; year 2008 and year 2012: Recycling centre: 0% (2008) and 100% (2012), Recycling bin: 100% (2008) and 50% (2012), Recycling Van: 600% (2008) and 0% (2012), Silverbox: 300% (2008) and 0% (2012)}
- Strategy 10: {Effect of awareness creation = low (1%); Increase recycling facilities in 2-phase, year 2008 and year 2012 as given in Strategy 9}
- Strategy 11: {Effect of awareness creation = relatively high (10%); Increase recycling facilities in 2-phase; year 2008 and year 2012 as given in Strategy 9}
- Strategy 12: {Effect of awareness creation = status quo; Increase of recycling facilities in 2012: Recycling centre: 150% , Recycling bin: 500% , Recycling Van: 800% , Silverbox: 800% }
- Strategy 13: {Effect of awareness creation = relatively high (10%); Increase recycling facilities in 2012 by percentages given in Strategy 12}
- Strategy 14: {Effect of awareness creation = high (20%); Increase recycling facilities in 2012 by percentages given in Strategy 12}

3.2 Criteria for Consideration

We considered 5 criteria for evaluation:

1. Recycling rate (c_1): The main objective is to achieve the government recycling target of 22% by 2020
2. Construction cost (c_2): Recycling centre (from MYR 70,000); Recycling bins (from MYR 1500); Recycling Van (from MYR 70,000); Silverboxes (approximately MYR 5,000)
3. Operating/ maintenance cost (c_3): Operating/ maintenance cost is given using score from 1 to 5. Score '1' indicates high maintenance and score '5' indicates low maintenance. Higher score is desirable.
4. Social impact (employment, economic benefits to the residents)(c_4): This criterion is described using scores 1 to 5. Score '1' indicates low impact, whilst score '5' indicates high impact. Higher score is desirable.
5. Convenience for the residents (accessibility)(c_5): We use scores 1 to 5, with score '1' indicating low accessibility and score '5' indicating high accessibility. Higher score is desirable.

3.3 Construction of the Outranking Relation

PROMETHEE assumes that the decision-maker is able to weigh the criteria appropriately, at least when the number of criteria is not too large. This may be difficult to achieve by an inexperienced user. Macharis [4] suggested utilizing Analytic Hierarchy Process (AHP)(see, [6]). We constructed the reciprocal matrix and derived the priority vector: (0.470, 0.144, 0.079, 0.045, 0.262). The following table presents a summary of criteria and strategies:

	<i>c</i> ₁	<i>c</i> ₂	<i>c</i> ₃	<i>c</i> ₄	<i>c</i> ₅		<i>c</i> ₁	<i>c</i> ₂	<i>c</i> ₃	<i>c</i> ₄	<i>c</i> ₅
<i>a</i> ₁	0.58	100000	5	1	2	<i>a</i> ₈	0.86	230500	4	2	2
<i>a</i> ₂	0.68	150000	4	1	2	<i>a</i> ₉	1.30	911500	3	3	3
<i>a</i> ₃	2.81	150000	3	1	2	<i>a</i> ₁₀	1.52	961500	3	3	3
<i>a</i> ₄	5.98	150000	3	2	2	<i>a</i> ₁₁	5.99	961500	3	3	3
<i>a</i> ₅	12.15	150000	2	2	2	<i>a</i> ₁₂	1.87	1592500	2	4	4
<i>a</i> ₆	22.69	150000	2	3	2	<i>a</i> ₁₃	8.45	1642500	1	5	5
<i>a</i> ₇	37.18	150000	2	3	2	<i>a</i> ₁₄	29.7	1642500	1	5	5
weight							0.470	0.144	0.079	0.045	0.262

Let threshold values *p*_{*i*}(preference), *q*_{*i*}(indifference) and *v*_{*i*}(veto) of each criterion *c*_{*i*} be: *p*₁ = 3, *q*₁ = 0.5, *v*₁ = 10, *p*₂ = 200,000, *q*₂ = 100,000, *v*₂ = 10,000,000, *p*₃ = 3, *q*₃ = 1, *v*₃ = 5 for *i* = 3,4,5. We constructed outranking relation matrix.

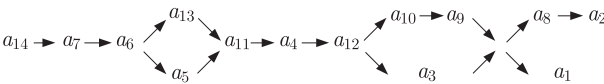
$$\left(\begin{array}{cccccccccccccccc} 1 & 1 & 0.675 & 0.53 & \epsilon & \epsilon & \epsilon & 0.956 & 0.792 & 0.751 & 0.364 & 0.531 & 0.013 & \epsilon \\ 1 & 1 & 0.694 & 0.53 & \epsilon & \epsilon & \epsilon & 1 & 0.811 & 0.770 & 0.364 & 0.550 & 0.014 & \epsilon \\ 0.961 & 1 & 1 & 0.53 & 0.106 & \epsilon & \epsilon & 1 & 0.834 & 0.834 & 0.364 & 0.68 & 0.027 & \epsilon \\ 0.961 & 1 & 1 & 1 & 0.53 & \epsilon & \epsilon & 1 & 0.856 & 0.856 & 0.856 & 0.703 & 0.179 & \epsilon \\ 0.921 & 0.961 & 1 & 1 & 1 & \epsilon & \epsilon & 0.961 & 0.856 & 0.856 & 0.856 & 0.703 & 0.549 & \epsilon \\ 0.921 & 0.961 & 1 & 1 & 1 & 1 & \epsilon & 0.961 & 0.856 & 0.856 & 0.856 & 0.725 & 0.572 & 0.047 \\ 0.921 & 0.961 & 1 & 1 & 1 & 1 & 1 & 0.961 & 0.856 & 0.856 & 0.856 & 0.725 & 0.572 & 0.572 \\ 1 & 1 & 0.727 & 0.53 & \epsilon & \epsilon & \epsilon & 1 & 0.856 & 0.826 & 0.386 & 0.607 & 0.028 & \epsilon \\ 0.961 & 1 & 0.810 & 0.53 & \epsilon & \epsilon & \epsilon & 1 & 1 & 1 & 0.53 & 0.843 & 0.123 & \epsilon \\ 0.961 & 1 & 0.851 & 0.53 & \epsilon & \epsilon & \epsilon & 1 & 1 & 1 & 0.53 & 0.856 & 0.133 & \epsilon \\ 0.961 & 1 & 1 & 1 & 0.53 & \epsilon & \epsilon & 1 & 1 & 1 & 1 & 0.856 & 0.334 & \epsilon \\ 0.921 & 0.961 & 0.917 & 0.53 & \epsilon & \epsilon & \epsilon & 0.961 & 1 & 1 & 0.53 & 1 & 0.53 & \epsilon \\ 0.921 & 0.921 & 0.961 & 0.961 & 0.53 & \epsilon & \epsilon & 0.921 & 0.961 & 0.961 & 0.961 & 1 & 1 & \epsilon \\ 0.921 & 0.921 & 0.961 & 0.961 & 1 & 1 & 0.406 & 0.921 & 0.961 & 0.961 & 0.961 & 1 & 1 & 1 \end{array} \right)$$

where $\epsilon = 0.0001$. From the eigenvector procedure using weighted preference flows, we have:

$\psi^+ = (0.0479, 0.0487, 0.0535, 0.0688, 0.0766, 0.0894, 0.1119, 0.0505, 0.0576, 0.0581, 0.0753, 0.0631, 0.0806, 0.1180)$

$\psi^- = (0.1260, 0.1290, 0.1086, 0.0798, 0.0150, 0.0000, 0.0000, 0.1283, 0.1166, 0.1147, 0.0678, 0.0941, 0.0201, 0.0000)$

$\lambda_{max} = 7.706$ and thus the final ranking :



From the study, Strategy 14 were ranked first, followed by Strategies 7 and 6, even though construction cost and maintenance cost is highest. This is due to the fact that recycling facilities were increased and awareness creation was high rate of 20%. Awareness creation under Strategy 7 is 30% and Strategy 6 is 25%. Even without increasing the recycling facilities, recycling rate can be increased towards target of 2020 but the effect of awareness creation must be at least 20% or higher. The recycling rate is higher when increase in recycling facilities is complemented by awareness creation. Strategy 1 and 2 is not favored because the effect of awareness creation is nil or low. Strat-

egy 8 and 9 are less favored due to the fact that increase in the number of recycling facilities is proposed without taking into consideration awareness creation. Strategy 11 is more favored compared to Strategy 8 and 9 because maximum increase in recycling facilities were considered. Even though the cost involved will be higher, the social impact on the residents will be higher too. Furthermore, with more recycling facilities distributed all over Subang Jaya, the facilities are easily accessible.

4 Concluding Remarks

We have demonstrated a new exploitation procedure based on the weighted preference in-flows and out-flows in the outranking analysis. The results from the outranking analysis using eigenvector clearly states that in order to increase recycling rate to 22% by 2020, MPSJ should focus not only on increasing the facilities but on awareness creation too. In fact, awareness creation should be given utmost attention in order to create the civic-mindedness among the residents. It can be seen that awareness creation on itself can increase the recycling rate of MPSJ. However, awareness without accessibility to the facilities is not a rational thought.

References

1. Brans, J.P., Vincke, Ph., 1985, A preference ranking organization method: The PROMETHEE method, *Management Science*, **31**, 647-656.
2. Chenayah, S., Takeda, E., 2006, A generalized Procedure in PROMETHEE Analysis, with an application to the evaluation of Recycling Strategies in Malaysia, *Asia Pacific Management Review*, **11** (1), 85-96.
3. Fodor, J., Roubens, M., 1994, *Fuzzy Preference Modelling and Multicriteria Decision Support*, Kluwer Academic Publishers.
4. Macharis, C., Springael, J., Brucker, K., and Verbeke, A., 2004. PROMETHEE and AHP: The design of operational synergies in multicriteria analysis. Strengthening PROMETHEE with ideas of AHP, *European Journal of Operational Research*, **153**, 307-317.
5. Roy, B., 1996, *Multicriteria Methodology for Decision Aiding*, Kluwer Academic Publishers, Dordrecht, Boston, London.
6. Saaty, T.L., 1980, *The Analytic Hierarchy Process*, McGraw-hill, New York.

Dimensionality Reduction in Multiobjective Optimization: The Minimum Objective Subset Problem

Dimo Brockhoff and Eckart Zitzler

Computer Engineering and Networks Laboratory, ETH Zurich, Switzerland
brockhoff@tik.ee.ethz.ch
zitzler@tik.ee.ethz.ch

Summary. The number of objectives in a multiobjective optimization problem strongly influences both the performance of generating methods and the decision making process in general. On the one hand, with more objectives, more incomparable solutions can arise, the number of which affects the generating method's performance. On the other hand, the more objectives are involved the more complex is the choice of an appropriate solution for a (human) decision maker. In this context, the question arises whether all objectives are actually necessary and whether some of the objectives may be omitted; this question in turn is closely linked to the fundamental issue of conflicting and non-conflicting optimization criteria. Besides a general definition of conflicts between objective sets, we here introduce the $\mathcal{NP}$ -hard problem of computing a minimum subset of objectives without losing information (MOSS). Furthermore, we present for MOSS both an approximation algorithm with optimum approximation ratio and an exact algorithm which works well for small input instances. We conclude with experimental results for a random problem and the multiobjective 0/1-knapsack problem.

1 Motivation

It is indisputable, that a high number of criteria in a multiobjective optimization problem [5, 10] can cause additional difficulties compared to a low-dimensional problem. The number of incomparable solutions can raise [7], the (human) decision making process becomes harder and generating methods may require substantially more computational resources¹ when more criteria are involved in the problem formulation. Consequently, the question arises whether it is possible to omit some of the objectives without changing the characteristics of the underlying problem. Furthermore, one may ask under which conditions such an objective reduction is feasible and how a minimum set of objectives can be computed.

¹ For example, when based on the $\mathcal{S}$ -metric [15].

These questions have gained only little attention in the literature so far. Transferred to the multiobjective optimization setting, closely related research topics as principal component analysis [9] or dimension theory [14] aim at determining a (minimum) set of *arbitrary* objective functions that preserves (most of) the problem characteristics; however, here we are interested in determining a minimum subset of *original* objectives that maintains the order on the search space. Known attempts dealing with the latter do not preserve the dominance structure [4] or are not suitable for black-box scenarios [1]. Furthermore, there are a few studies that investigate the relationships between objectives in terms of conflicting and nonconflicting optimization criteria. The definitions of conflicting objectives are based on trade-offs within the Pareto optimal set [1, 3, 8] or on the number of incomparable solutions [11, 13]. Conflicts are either defined between objective pairs [1, 8, 11, 13] or as a property of the entire objective set [3]. However, these definitions are not sufficient to indicate whether objectives can be omitted or not; examples can be constructed, cf. [2], where all objectives are conflicting according to [3, 11, 13], but one among the given objectives can be removed while preserving the search space order.

Assuming that the decision making process is postponed after the search process, this paper addresses two open issues for a given set of (trade-off) solutions: (i) deriving general conditions under which certain objectives may be omitted and (ii) computing a minimum subset of objectives needed to preserve the problem structure between the given solutions. In particular, we

- propose a generalized notion of objective conflicts which comprises the definitions of Deb [3], Tan et al. [13], and Purshouse and Fleming [11],
- specify on this basis a necessary and sufficient condition under which objectives can be omitted,
- introduce the $\mathcal{NP}$ -hard problem of minimum objective subsets (MOSS),
- provide both an approximation algorithm and an exact algorithm for MOSS,
- validate our approach in additional experiments by comparing the algorithms and investigating the influence of the number of objectives and the search space size.

2 A Notion of Objective Conflicts

Without loss of generality, in this paper we consider a minimization problem with k objective functions $f_i : X \rightarrow \mathbb{R}$, $1 \leq i \leq k$, where the vector function $f := (f_1, \dots, f_k)$ maps each solution $\mathbf{x} \in X$ to an objective vector $f(\mathbf{x}) \in \mathbb{R}^k$. Furthermore, we assume that the underlying dominance structure is given by the weak Pareto dominance relation² which is defined as follows: $\preceq_{\mathcal{F}} :=$

² Any other preorder rel like the ε -dominance relation can be used as well if $rel_{\mathcal{F}} = \bigcap_{1 \leq i \leq k} rel_i$ holds for any set $\mathcal{F} = \{f_1, \dots, f_k\}$ of k objective functions. The proof of the equivalence is simple for $\preceq_{\mathcal{F}}$ and can be found in [2].

$\{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x}, \mathbf{y} \in X \wedge \forall f_i \in \mathcal{F}' : f_i(\mathbf{x}) \leq f_i(\mathbf{y})\}$, where $\mathcal{F}'$ is a set of objectives with $\mathcal{F}' \subseteq \mathcal{F} := \{f_1, \dots, f_k\}$. For better readability, we will sometimes only consider the objective functions' indices, e.g., $\mathcal{F}' = \{1, 2, 3\}$ instead of $\mathcal{F}' = \{f_1, f_2, f_3\}$. The Pareto (optimal) set is given as $\{\mathbf{x} \in X \mid \nexists \mathbf{y} \in X \setminus \{\mathbf{x}\} : \mathbf{y} \preceq_{\mathcal{F}} \mathbf{x} \wedge \mathbf{x} \not\preceq_{\mathcal{F}} \mathbf{y}\}$.

In the following, we introduce a general concept of conflicts between sets of objectives and formalize the notion of redundant objective sets. On this basis, we then propose two algorithms to exactly resp. approximately compute a minimum set of objectives, which induces the same preorder on X as the whole set of objectives.

Definition 1. Let $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{F}$ be two sets of objectives. We call

- $\mathcal{F}_1$ nonconflicting with $\mathcal{F}_2$ iff $\preceq_{\mathcal{F}_1} \subseteq \preceq_{\mathcal{F}_2} \wedge \preceq_{\mathcal{F}_2} \subseteq \preceq_{\mathcal{F}_1}$
- $\mathcal{F}_1$ weakly conflicting with $\mathcal{F}_2$ iff $(\preceq_{\mathcal{F}_1} \subseteq \preceq_{\mathcal{F}_2} \wedge \preceq_{\mathcal{F}_2} \not\subseteq \preceq_{\mathcal{F}_1})$ or $(\preceq_{\mathcal{F}_2} \subseteq \preceq_{\mathcal{F}_1} \wedge \preceq_{\mathcal{F}_1} \not\subseteq \preceq_{\mathcal{F}_2})$
- $\mathcal{F}_1$ strongly conflicting with $\mathcal{F}_2$ iff $\preceq_{\mathcal{F}_1} \not\subseteq \preceq_{\mathcal{F}_2} \wedge \preceq_{\mathcal{F}_2} \not\subseteq \preceq_{\mathcal{F}_1}$

Two sets of objectives $\mathcal{F}_1, \mathcal{F}_2$ are called nonconflicting if and only if the corresponding relations $\preceq_{\mathcal{F}_1}$ and $\preceq_{\mathcal{F}_2}$ are identical but not necessarily $\mathcal{F}_1 = \mathcal{F}_2$. If $\mathcal{F}_1 \subset \mathcal{F}_2$ and $\mathcal{F}_1$ is nonconflicting with $\mathcal{F}_2$ we can simply omit all objectives in $\mathcal{F}_2 \setminus \mathcal{F}_1$ without influencing the preorder on X . This formulation of conflicting objectives can be regarded as a generalization of the conflict definition in [11]; moreover, it also comprises the notions in [3] and [13], see [2] for details.

A *minimal objective set w. r. t. $\mathcal{F}$* is a subset $\mathcal{F}'$ of the original objectives $\mathcal{F}$ that cannot be further reduced without changing the associated preorder, i.e., $\preceq_{\mathcal{F}'} = \preceq_{\mathcal{F}}$ but $\nexists \mathcal{F}'' \subset \mathcal{F}' : \preceq_{\mathcal{F}''} = \preceq_{\mathcal{F}'}$. A *minimum objective set w. r. t. $\mathcal{F}$* is the smallest possible subset of the original objectives $\mathcal{F}$ that preserves the original order on the search space. By definition, every minimum objective set is minimal, but not all minimal sets are at the same time minimum.

Definition 2. A set $\mathcal{F}$ of objectives is called *redundant* if and only if there exists $\mathcal{F}' \subset \mathcal{F}$ that is *minimal w. r. t. $\mathcal{F}$* .

This definition of redundancy represents a necessary and sufficient condition for the omission of objectives.

3 The Minimum Objective Subset Problem

Given a multiobjective optimization problem with the set $\mathcal{F}$ of objectives, the question arises whether objectives can be omitted without changing the order on the search space. If there is an objective subset $\mathcal{F}' \subseteq \mathcal{F}$ such that $\mathbf{x} \preceq_{\mathcal{F}'} \mathbf{y}$ holds for all solutions $\mathbf{x}, \mathbf{y} \in X$ if and only if $\mathbf{x} \preceq_{\mathcal{F}} \mathbf{y}$, we can omit all objectives in $\mathcal{F} \setminus \mathcal{F}'$ while preserving the preorder on X . Concerning the last section, we are interested in identifying a minimum objective subset with respect to $\mathcal{F}$, yielding a slighter representation of the same multiobjective

optimization problem. Formally, this problem, the $\mathcal{NP}$ -hardness of which can easily be shown [2], is stated as follows.

Definition 3. *Given a multiobjective optimization problem with objective set $\mathcal{F} = \{f_1, \dots, f_k\}$, the problem **MINIMUM OBJECTIVE SUBSET (MOSS)** is defined as follows.*

Instance: A set $A \subseteq X$ of solutions, the generalized weak Pareto dominance relation $\preceq_{\mathcal{F}}$ and for all objective functions $f_i \in \mathcal{F}$ the single relations $\preceq_i$ where $\bigcap_{1 \leq i \leq k} \preceq_i = \preceq_{\mathcal{F}}$.

Task: Compute an objective subset $\mathcal{F}' \subseteq \mathcal{F}$ of minimum size with $\preceq_{\mathcal{F}'} = \preceq_{\mathcal{F}}$.

Note, that the dominance relations in Def. 3 have to be known only on the set $A \subseteq X$ of solutions, which, in practice, will mostly be rather a Pareto set approximation than the entire search space.

Algorithm 1 A greedy algorithm for MOSS

```

 $E := \preceq_{\mathcal{F}}^C$  where  $\preceq_{\mathcal{F}}^C := (A \times A) \setminus \preceq_{\mathcal{F}}$ 
 $I := \emptyset$ 
while  $E \neq \emptyset$  do
    choose an  $i \in (\{1, \dots, k\} \setminus I)$  such that  $|\preceq_i^C \cap E|$  is maximal
     $E := E \setminus \preceq_i^C$ 
     $I := I \cup \{i\}$ 
end while

```

Algorithm 2 An exact algorithm for MOSS

```

 $S := \emptyset$ 
for each pair  $\mathbf{x}, \mathbf{y} \in A$  of solutions do
     $S_x := \{\{i\} \mid i \in \{1, \dots, k\} \wedge \mathbf{x} \preceq_i \mathbf{y} \wedge \mathbf{y} \not\preceq_i \mathbf{x}\}$ 
     $S_y := \{\{i\} \mid i \in \{1, \dots, k\} \wedge \mathbf{y} \preceq_i \mathbf{x} \wedge \mathbf{x} \not\preceq_i \mathbf{y}\}$ 
     $S_{xy} := S_x \sqcup S_y$  where
         $S_1 \sqcup S_2 := \{s_1 \cup s_2 \mid s_1 \in S_1 \wedge s_2 \in S_2 \wedge (\nexists p_1 \in S_1, p_2 \in S_2 : p_1 \cup p_2 \subset s_1 \cup s_2)\}$ 
    if  $S_{xy} = \emptyset$  then  $S_{xy} := \{1, \dots, k\}$ 
     $S := S \sqcup S_{xy}$ 
end for
Output a smallest set  $s_{\min}$  in  $S$ 

```

With Algorithm 1, we present an approximation algorithm with polynomial running time for the MOSS problem, the approximation ratio of which is optimal. Known results for the set cover problem [6, 12] can be used to prove the algorithm's $\Theta(\log |A|)$ approximation ratio [2]³. As an exact algorithm for

³ For the proof of correctness and the $O(k \cdot |A|^2)$ runtime, we also refer to [2].

MOSS, we present Algorithm 2, the running time of which is polynomial in the size of A but exponential in the number of objectives⁴. That no correct predication is possible whether a set of objectives is redundant, by observing only relations between objective subsets of constant size, can be likewise derived from the $\mathcal{NP}$ -hardness of the MOSS problem. Thus, Algorithm 2 is forced to examine the type of conflict between all possible objective subsets. Note that Algorithm 2 computes all minimal objective subsets for a given MOSS instance. For further details on both algorithms, we refer to [2].

4 Experiments

The following experiments serve two goals: (i) to investigate the size of a minimum objective subset depending on the size of the search space and the number of original objective functions, and (ii) to compare the two proposed algorithms with respect to the size of the generated objective subsets and their running times. These issues have been considered both on the whole search space X for a random problem and on Pareto set approximations $A \subseteq X$ of the multiobjective 0/1-knapsack problem. For details on the experiment settings, we refer to [2] due to space limitations.

In a first experiment, we generated the objective values for a set of solutions X at random, where for each combination of search space size $|X|$ and number of objectives k , 100 independent samples were considered. Figure 2 shows the sizes of the minimum objective subsets, computed with the exact algorithm, Fig. 1 the results for both algorithms on a fixed problem size. For

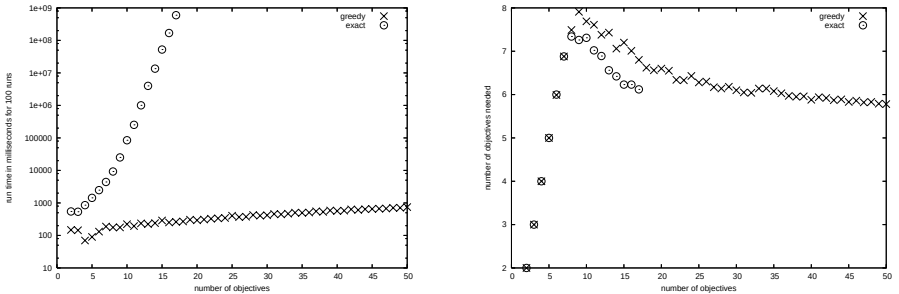


Fig. 1. Comparison between the greedy and the exact algorithm for the random problem with 32 solutions on a linux computer (SunFireV60x with 3060 MHz) over 100 runs: summed running times (*left*); averaged sizes of the computed minimum / minimal sets (*right*)

a second experiment, both the exact and the approximation algorithm were applied to Pareto set approximations A for the knapsack problem, generated

⁴ The runtime bounds of $O(|A|^2 \cdot k \cdot 2^k)$, resp. $\Omega(|A|^2 \cdot 2^{k/3})$, are proved in [2].

by an evolutionary algorithm. Figure 3 shows the results together with the algorithms’ runtimes.

Two main observations can be made. First, the experiments, based both on the entire decision space and on Pareto set approximations, show that the omission of objectives without information loss is possible. Furthermore, the minimum number of objectives decreases the more objectives are involved in the problem formulation, and the larger the search space the more objectives are in a minimum objective set. Second, in comparison to the exact algorithm, the greedy algorithm shows nearly the same output quality regarding the computed objective set size but is much faster. Due to the sizes of the computed subsets which are—in all of our experiments—less than one objective away from the optimum, the greedy algorithm seems to be applicable for more complex problems, particularly by virtue of its small running time.

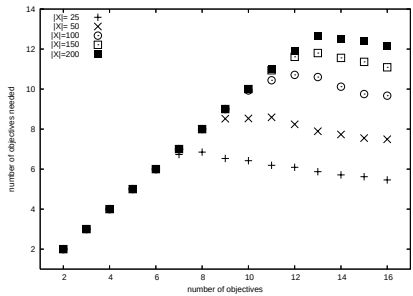


Fig. 2. Random model: The size of a minimum subset, computed with the exact algorithm, versus the number of objectives in the problem formulation

<i>k</i>	objective set size		runtime [ms]	
	greedy	exact	greedy	exact
5	4	4	47	196
10	5	5	46	2,271
15	8	8	67	87,113
20	13	13	88	90,524
25	16	16	78	$\approx 2.5 \cdot 10^6$
30	14	13	87	$\approx 15 \cdot 10^6$

Fig. 3. Results for both algorithms on a Pareto set approximation of the knapsack problem with different number *k* of objectives

5 Conclusions

From a practical point of view, the present study provides a first step towards dimensionality reduction of the objective space in multiple criteria optimization scenarios. We have introduced the minimum objective subset problem (MOSS) and proposed both an exact and a greedy algorithm for MOSS. The presented algorithms can be particularly useful to analyze Pareto set approximations generated by search procedures. How to integrate dimensionality reduction methods into search heuristics and the question whether this makes sense in general, has to be disputed in future work.

References

1. P. J. Agrell. On redundancy in multi criteria decision making. *European Journal of Operational Research*, 98(3):571–586, 1997.
2. D. Brockhoff and E. Zitzler. On Objective Conflicts and Objective Reduction in Multiple Criteria Optimization. TIK Report 243, February 2006.
3. K. Deb. *Multi-objective optimization using evolutionary algorithms*. Wiley, Chichester, UK, 2001.
4. K. Deb and D. K. Saxena. On finding pareto-optimal solutions through dimensionality reduction for certain large-dimensional multi-objective optimization problems. Kangal report no. 2005011, December 2005.
5. M. Ehrgott. *Multicriteria Optimization*. Springer Berlin Heidelberg, 2005.
6. U. Feige. A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45(4):634–652, 1998.
7. C. M. Fonseca and P. J. Fleming. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation*, 3(1):1–16, 1995.
8. T. Gal and H. Leberling. Redundant objective functions in linear vector maximum problems and their determination. *European Journal of Operational Research*, 1(3):176–184, 1977.
9. I. T. Jolliffe. *Principal component analysis*. Springer, 2002.
10. K. M. Miettinen. *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.
11. R. C. Purshouse and P. J. Fleming. Conflict, harmony, and independence: Relationships in evolutionary multi-criterion optimisation. In *EMO 2003 Proceedings*, pages 16–30. Springer, Berlin, 2003.
12. P. Slavík. A tight analysis of the greedy algorithm for set cover. In *STOC '96: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 435–441, New York, NY, USA, 1996. ACM Press.
13. K. C. Tan, E. F. Khor, and T. H. Lee. *Multiobjective Evolutionary Algorithms and Applications*. Springer, London, 2005.
14. W. T. Trotter. *Combinatorics and Partially Ordered Sets: Dimension Theory*. The Johns Hopkins University Press, Baltimore and London, 1992.
15. L. While. A new analysis of the lebmeasure algorithm for calculating hypervolume. In *EMO 2005 Proceedings*, pages 326–340. Springer, 2005.

Multikriterielle Entscheidungsunterstützung zur Auswahl von Lagersystemen in der Ersatzteillogistik

Gerrit Reiniger and Martin Josef Geiger

Lehrstuhl für Industriebetriebslehre, Universität Hohenheim, Germany
reinigerrit@web.de
mjgeiger@uni-hohenheim.de

Summary. Der Grundstein für eine erfolgreiche Ersatzteil-Lagerlogistik wird mit der Entscheidung über die passenden Lagersysteme gelegt. In der Praxis konkurrieren dabei multiple Zielsetzungen wie z. B. ein zu maximierender Automatisierungsgrad mit den resultierenden zu minimierenden Investitionen. Der Beitrag stellt ein multikriterielles Entscheidungsmodell vor, das speziell im Ersatzteilbereich die Auswahlproblematik anhand von Kennzahlen strukturiert und Möglichkeiten zur Entscheidungsunterstützung anbietet. Zum Einsatz kommt ein dreistufiges Modell auf Grundlage des Analytic Hierarchy Process, das mit Hilfe der Software Expert Choice die Eignung verschiedener Alternativen in unterschiedlichen Lagerzonen ermitteln kann. In Abhängigkeit der zugrunde liegenden Teilestruktur des betrachteten Unternehmens werden zunächst in Frage kommende Lagersysteme ausgewählt und anschließend mit Hilfe des Modells bewertet. Eine Sensitivitätsanalyse sichert die Stabilität des Ergebnisses ab.

1 Einleitung

Die Erwartungen der Kunden an Serviceleistungen im After-Sales Bereich sind besonders bei hochwertigen Produkten sehr hoch. Ein Ersatzteilbedarf ist fast immer damit verbunden, dass der Kunde vor dem Ersatzteilkauf ein Negativerlebnis mit dem Produkt verbindet, Ersatzteile sind also auch häufig Teil einer Kundendienstleistung. Um in diesem Bereich einen Wettbewerbsvorteil gegenüber Konkurrenten durchsetzen zu können, muss durch exzellente Ersatzteillogistik eine hoher Service- und Liefergrad erreicht werden. Nur eine schnelle Kundendienstabwicklung kann langfristig Kunden an eine Marke binden und die Wettbewerbsfähigkeit gegenüber Konkurrenten erhalten.

Der Grundstein für eine erfolgreiche Ersatzteil-Lagerlogistik wird mit der Entscheidung über die passenden Lagersysteme gelegt. Um diesen Entscheidungsprozess transparent zu machen, entwickelt der vorliegende Artikel ein Entscheidungsmodell, welches an individuelle Entscheidungssituationen ange-

passt werden kann. Es soll in der Lage sein, den Entscheider bei der Auswahl von Lagersystemen zu unterstützen, indem grundsätzlich in Frage kommende Alternativen ausgewählt und anhand von Kennzahlen bewertet werden. Die Bedeutung, die der Entscheider den jeweiligen Kennzahlen zumisst und die Eignung der Lagersysteme in Bezug auf die Kennzahlen können individuell festgelegt werden.

2 Systeme der Ersatzteillogistik

2.1 Bodenlagerung

In der einfachsten Form der Lagerung verzichtet man auf Regale. Das Lagergut wird mit oder ohne Ladehilfsmittel artikelrein im direkten Zugriff auf dem Boden gelagert und nicht gestapelt [3].

Stapelfähiges Lagergut oder Güter in Ladehilfsmitteln können in Form einer Blocklagerung übereinander und gegebenenfalls hintereinander gelagert werden. Eine Bodenblocklagerung kommt für die Lagerung von großen Beständen pro Artikel bei einem kleinen Sortiment unterschiedlicher Umschlagsleistung in Frage.

2.2 Statische Regallagerung

(Paletten-)Ein- und Durchfahrregalen liegt das Prinzip der Blocklagerung zu Grunde, wobei das Lagergut bei dieser Lagertechnik nicht stapelbar sein muss, da die Ladeeinheiten vom Regal getragen werden [5]. Geeignet ist diese Lagertechnik daher für die Lagerung von großen Mengen pro Artikel und evtl. von Gütern mit langer Verweildauer wie z. B. Saisonartikel, die ein hohes Gewicht haben, einen geringen Umschlag, nicht stapelbar sind und nur in einer relativ kleinen Anzahl unterschiedlicher Artikel vorhanden sind.

Die Lagerung von Teilen in Fachbodenregalen erfolgt auf geschlossenen Fachböden aus Holz oder Stahlblech in mehreren Ebenen übereinander und entspricht dem Prinzip der Zeilenlagerung. Fachbodenregale eignen sich sehr gut für die Lagerung von kleinen bis mittleren Mengen pro Artikel, wenn ein sehr breites Sortiment gegeben ist, was z. B. bei der Ersatzteillagerung häufig der Fall ist. Grundsätzlich ist die Lagerung unterschiedlichster Güter denkbar, besonders effizient ist die Lagertechnik aber für Kleinteile.

Palettenregale werden unterschieden in Palettenflachregallager (bis ca. sieben Meter Höhe), mittelhohe Palettenlager (sieben bis ca. 15 Meter Höhe) und Paletten-Hochregallager (über 15 Meter Höhe). Sämtliche Güter in Ladungsträgern können in dieser Form gelagert werden. Ein flaches Palettenregal findet in Lagern mit großer Artikelvielfalt Anwendung, in denen große Mengen mittelschwerer bis schwerer Artikel aufbewahrt werden. Ein hohes Palettenregal findet in Lagern mit großer Menge je Artikel, breitem Sortiment und mittlerer bis hoher Umschlagsleistung Anwendung.

Behälterregale heißen auch Kleinteile- oder Kompaktlager und sind dazu geeignet, Kleinteile in Behältern, Kisten oder auf Tablaren zu lagern. Ein breites Sortiment mit einer begrenzten Menge pro Artikel kann so für Kommissionierzwecke bereitgestellt werden.

2.3 Dynamische Regallagerung

Die Lagerung in Durchlauf- und Einschubregalen erfolgt sortenrein in Kanälen, die neben- und übereinander angeordnet sind. Innerhalb der Kanäle wird das Ladegut nach dem Schwerkraftprinzip oder dem Antriebsprinzip bewegt. Einschubregale werden nur von einer Seite be- und entladen, Durchlaufregale von beiden Seiten. Einsatzmöglichkeiten sind die Bereitstellung von großen Mengen pro Artikel bei kleiner bis mittlerer Artikelzahl und einer mittleren bis hohen Umschlagsleistung.

Die Fachböden eines horizontalen Umlaufregals oder Karussellregals werden manuell oder elektromechanisch horizontal in die Entnahmeposition gefahren. Wenn kein Zugriff erfolgt, ruht das Lagergut. Bis zu drei Umlaufeinheiten sind im direkten Zugriff. Einsatzgebiete finden Karussellregale vor allem für kleine bis mittlere Mengen pro Artikel mit mittlerem bis großem Sortiment, speziell bei Kleinteilelagerung.

Die Regalböden bei vertikalen Umlaufregalen laufen in vertikaler Richtung um, dieser Regaltyp wird daher auch als Paternosterregal bezeichnet. Im direkten Zugriff sind nur ein bis zwei Regalböden, der Antrieb erfolgt mit einem Elektromotor. Einsatzgebiete des Paternosters findet man bei kleinen bis mittleren Mengen pro Artikel bei mittlerem bis großem Sortiment und mittlerer Umschlagsleistung, sehr häufig im Kleinteilebereich, in Ersatzteillagern, Dokumentenlagern oder Werkzeuglagern.

Verschieberegale entstehen, wenn Regaltypen wie Fachbodenregale oder Palettenregale auf Unterwagen montiert werden, um sie auf Führungsschienen manuell oder mit Antrieb verschieben zu können. Nicht benötigte Regale bilden dann eine Blocklagerung mit hohem Flächennutzungsgrad, da nur dort eine Gasse gebildet wird, wo Teile entnommen werden sollen. Einsatzgebiete finden sich bei kleinen bis mittleren Mengen pro Artikel bei mittlerem bis großem Sortiment und geringer Umschlagshäufigkeit.

3 Ein multikriterielles Entscheidungsmodell

3.1 Der Analytic Hierarchy Process

Der Analytic Hierarchy Process (AHP) ist ein Verfahren zur Entscheidungsunterstützung, das bei multikriteriellen Zielsetzungen in einer Situation der Sicherheit zur Anwendung kommt [4]. Das Modell kann sowohl Intuition als auch Rationalität abbilden, um eine Alternativauswahl anhand von mehreren Kriterien zu bewerten. Der Entscheider vergleicht die Alternativen

immer paarweise. AHP lässt dabei inkonsistente Aussagen zu und hilft, diese zu verbessern. Die einfachste Form, ein Entscheidungsproblem zu strukturieren, ist eine dreistufige Hierarchie: Das Entscheidungsziel steht auf der ersten Ebene, gefolgt von der zweiten Ebene, auf der die Kriterien angeordnet sind und schließlich der dritten Ebene, auf der die Alternativen bewertet werden.

3.2 Formulierung und Anwendung eines Entscheidungsmodells

Im Hinblick auf die hohe Bedeutung der Zugriffshäufigkeit in Ersatzteillagern kann man nach einer ABC-Analyse verschiedene Szenarien bilden. Denkbar wäre neben einer Teilung in hoch- (A), mittel- (B) und niederfrequente (C) Teile auch eine Unterscheidung in mehr als drei Zonen. Um die Alternativen sinnvoll miteinander vergleichen zu können, sind die zu erwartenden Dimensionen jeder Lagerzone für jede Alternative vor der Bewertung zu ermitteln.

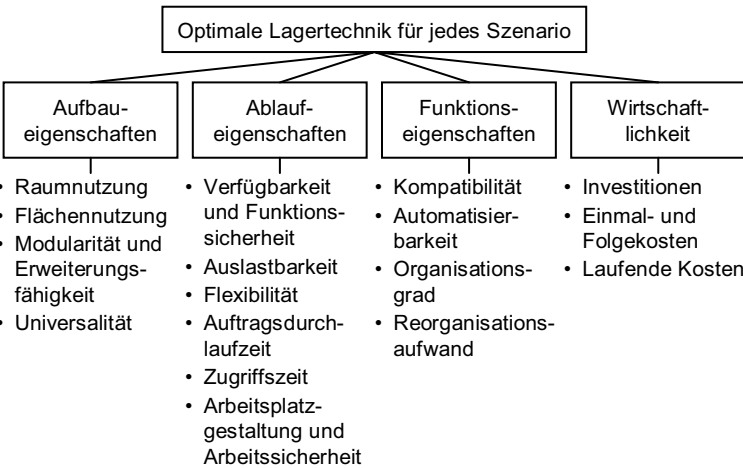


Fig. 1. Zielhierarchie für die Systemwahl

Mit den vier Kriterien Aufbau-, Ablauf-, Funktionseigenschaften und Wirtschaftlichkeit kann für jede Lagerzone genau festgelegt werden, welche Bedeutung ihnen in den einzelnen Zonen zugesprochen wird [2]. Abbildung 1 gibt die detaillierte Zielhierarchie der Kriterien an.

3.3 Berechnung und Kommentar der Ergebnisse

Das Ergebnis der Paarvergleiche für die Lagerung von hochfrequenten Teilen lautet wie in Tabelle 1 angegeben.

Die Ermittlung der Gewichtung ergab Werte für Aufbaueigenschaften in Höhe von 0,075, Ablaufeigenschaften 0,657, Funktionseigenschaften 0,051 und Wirtschaftlichkeit 0,217.

Table 1. Ergebnis für hochfrequente Teile

1. Fachbodenregal	0,312
2. Blocklagerung	0,278
3. Durchlaufregal	0,244
4. Automatisches Kleinteilelager	0,166

Über eine Sensitivitätsanalyse [1] wurde die Stabilität des Ergebnisses überprüft. Für Aufbaueigenschaften dominiert die Fachbodenanlage die Blocklagerung bis zu einer Gewichtung von $< 0,20$. Die Präferenz der Fachbodenanlage für Ablaufeigenschaften gilt bis zu einem Wert $> 0,38$, darunter wird die Blocklagerung präferiert, ab $0,26$ ein automatisches Kleinteilelager (AKL). Erst ab einer Gewichtung von $> 0,27$ der Funktionseigenschaften und $> 0,42$ der Wirtschaftlichkeit wird das Fachbodenregal vom automatischen Lager dominiert. Je weiter die Gewichtungspunkte von diesen Schwellenwerten entfernt sind, desto stabiler ist das Ergebnis.

Das Ergebnis der Paarvergleiche für die Lagerung von B-Teilen kann der folgenden Tabelle 2 entnommen werden.

Table 2. Ergebnis für mittelfrequente Teile

1. Automatisches Kleinteilelager	0,414
2. Fachbodenregal	0,226
3. Karussellregal	0,183
4. Paternoster	0,176

Die Gewichtung der Kriterien ergab Werte für die Aufbaueigenschaften von $0,125$, für Ablaufeigenschaften von $0,375$, für die Funktionseigenschaften von $0,375$ und für die Wirtschaftlichkeit von $0,125$. Eine abschließende Sensitivitätsanalyse konnte zeigen, dass auch in diesem Fall ein stabiles Ergebnis vorliegt.

Das Ergebnis der Paarvergleiche für die Lagerung von niederfrequenten (C) Teilen ist in Tabelle 3 angegeben.

Table 3. Ergebnis für niederfrequente Teile

1. Automatisches Kleinteilelager	0,537
2. Fachbodenregal	0,463

Aufbaueigenschaften erhielten ein Gewicht von $0,375$, Ablaufeigenschaften von $0,125$, Funktionseigenschaften von $0,125$ und die Wirtschaftlichkeit von $0,375$. Die Sensitivitätsanalyse zeigte trotz des sehr knappen Ergebnisses, dass

die Überlegenheit des AKL stabil ist und durch kleine Veränderungen der Variablen keine Rangumkehr ausgelöst wird.

4 Zusammenfassung und Fazit

In dieser Arbeit konnte ein hierarchisches Modell entwickelt werden, das die multikriterielle Entscheidungsproblematik bei der Auswahl von Lagersystemen abbildet. Trotz der in der Literatur sehr kontrovers diskutierten Methodik des AHP wird dieses Verfahren dem Komplexitätsgrad der Aufgabenstellung gerecht, ohne übertriebene Anforderungen an den Entscheidungsträger zu stellen. Einsatzmöglichkeiten des vorliegenden Modells ergeben sich, wenn man AHP als Instrument nutzt, um eine Tendenz der eigenen Prioritäten bei komplexen Problemen abzubilden. Die Lagerkennzahlen leisten dabei eine Hilfe, um auch Kriterien zu berücksichtigen, deren Bedeutung man vielleicht bisher unterschätzt hat.

Nach der vorliegenden Bewertung würde eine effiziente Lagerung der Teile in zwei Lagerzonen erfolgen: Die Fachbodenlagerung stellt die beste Möglichkeit dar, hochfrequente Teile in einer separaten Schnellläuferzone abzuwickeln. Dieses Ergebnis wurde durch eine Sensibilitätsanalyse bestätigt. Die Lagerung von mittel- und niederfrequenten Teilen soll dagegen nicht nochmals in verschiedene Zonen unterteilt werden, sondern gemeinsam in einem AKL erfolgen. Im vorliegenden Fall fiel die Entscheidung für mittelfrequente Teile recht deutlich auf das AKL, während in der Zone für niederfrequente Teile die Fachbodenlagerung und das AKL sehr nahe beieinander liegen. Das knappe Ergebnis in der Zone für niederfrequente Teile zeigt, dass das volle Potential eines AKL nur ausgeschöpft werden kann, wenn in den kommenden Jahren ein steigender Ersatzteilbedarf zu erwarten ist.

References

1. Belton V, Stewart TJ (2002) Multiple criteria decision analysis, an integrated approach, Kluwer Academic Publishers, Dordrecht
2. Dittrich M (2001) Lagerlogistik: neue Wege zur systematischen Planung, Hanser, München Wien
3. Jünemann R, Schmidt T (2000) Materialflusssysteme: Systemtechnische Grundlagen, Springer Verlag, Berlin Heidelberg New York
4. Saaty TL (1980) The Analytic Hierarchy Process, McGraw-Hill, New York
5. Schulte C (2005) Logistik: Wege zur Optimierung der Supply Chain, Vahlen, München

Network Optimization, Graphs and Traffic

Automatic Determination of Clusters

Bettina Hoser and Jan Schröder

Chair for Information Services and Electronic Markets, Universität Karlsruhe(TH), Germany

Bettina.Hoser@em.uni-karlsruhe.de

Jan.Schroeder@em.uni-karlsruhe.de

Summary. In this paper we propose an automatic method for spectral clustering of weighted directed graphs. It is based on the eigensystem of a complex Hermitian adjacency matrix $H_{n \times n}$. The number of relevant clusters is determined automatically. Nodes are assigned to clusters using the inner product matrix $S_{n \times n}$ calculated from a matrix $R_{n \times l}$ of the l eigenvectors as column vectors which correspond to the positive eigenvalues of H . It can be shown that by assigning the vertices of the network to clusters such that a node i belongs to cluster p_c if $Re(S_{i,p_c}) = \max_j Re(S_{i,j})$ an good partitioning can be found. Simulation results are presented.

1 Spectral Clustering Method

Spectral clustering of a graph by means of the eigensystem has a long tradition in graph theory. In 1970 Fiedler showed, that the the second smallest eigenvalue λ_{n-1} of a Laplace matrix depicts a measure for the algebraic connectivity of a graph [6]. Within the field of spectral clustering since then several algorithms were proposed that can be distinguished by the determination of the partitions. Some derive the partitions by doing recursive bipartitioning (e.g. [5, 8]), others directly derive the partitions by predefining the number of clusters (e.g. [4, 9, 1]) and Zien, Chan and Schlag [7] have proposed a hybrid method. The crucial point is always the definition of the initial clusters centers and their number. The second relevant step is the assignment of vertices to clusters. Especially for large graphs the predefinition of the number of clusters is difficult. If the partition into k clusters (k -partition) is derived directly, it is a question whether to use k eigenvectors which Chan et al. as well as Ng et al. [4, 9] did, or as Alpert and Yao suggested [1] to use many or even all eigenvectors.

We use an approach in between by taking the eigenvectors corresponding to all positive eigenvalues of the eigensystem (see Section 4). The spectral decomposition algorithm was proposed by Hoser and Geyer-Schulz [2] which is used to find structural details of the given directed and weighted network. The eigensystem of the graph is subsequently used to determine the clusters. Compared to many algorithms in the spectral partitioning field which work on undirected graphs like [4, 9, 8, 5] this algorithm and therefore the clustering works also on directed, weighted graphs.

2 Notations and Definitions

First we introduce the notation and some basic facts about complex numbers, Hilbert space and eigensystems of Hermitian matrices.

A complex number z can be represented in algebraic form or equivalently in exponential form as $z = a + ib = |z|e^{i\phi}$ with the real part of z being denoted as $Re(z) = a$, the imaginary part as $Im(z) = b$, the absolute value as $|z| = \sqrt{a^2 + b^2}$, and the phase as $0 \leq \phi = \arccos \frac{Re(z)}{|z|} \leq \pi$, with i as the imaginary unit ($i^2 = -1$). $\bar{z} = a - ib$ denotes the complex conjugate of z . Vectors will be complex valued column vectors unless otherwise stated.

Hilbert space is a complete normed inner product space. With the inner product of $\mathbf{x}, \mathbf{y} \in \Gamma^n$ being defined as $\langle \mathbf{x} | \mathbf{y} \rangle = \mathbf{x}^* \mathbf{y} = \sum_{k=1}^n \bar{x}_k y_k$, the norm is given defined as in Equation(1):

$$\sqrt{\langle \mathbf{x} | \mathbf{x} \rangle} = \|\mathbf{x}\| \quad (1)$$

From this it is clear that to minimize the distance between two points, characterized by vectors $\mathbf{x}, \mathbf{y}$ in an n -dimensional vector space is equivalent to maximizing the real part of the inner product of both. Let furthermore $\mathbf{x}, \mathbf{y}$ be normalized as in Equation(2),

$$\|\mathbf{x}\| = \|\mathbf{y}\| = 1 \quad (2)$$

then the distance between the two vectors is zero if the real part is equal to one, which is the maximum of the inner product (Equation(3)) between the vectors $\mathbf{x}$ and $\mathbf{y}$

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|^2 &= \langle \mathbf{x} - \mathbf{y} | \mathbf{x} - \mathbf{y} \rangle \\ &= \langle \mathbf{x} | \mathbf{x} \rangle + \langle \mathbf{y} | \mathbf{y} \rangle - \langle \mathbf{x} | \mathbf{y} \rangle - \langle \mathbf{y} | \mathbf{x} \rangle \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \langle \mathbf{x} | \mathbf{y} \rangle - \overline{\langle \mathbf{x} | \mathbf{y} \rangle} \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2Re(\langle \mathbf{x} | \mathbf{y} \rangle) \\ &= 2 - 2Re(\langle \mathbf{x} | \mathbf{y} \rangle). \end{aligned} \quad (3)$$

A matrix H is called Hermitian, if and only if $H^* = H$ with H^* representing the conjugate complex transpose of H . Hermitian matrices are also normal $HH^* = H^*H$. All eigenvalues of a Hermitian matrix are real and all eigenvectors are orthogonal and can be chosen as to define a complete orthonormal basis. Thus a matrix H can be written as the weighted sum of eigenprojectors $H = \sum_{k=1}^n \lambda_k P_k$ with λ_k representing the eigenvalues and $P_k = \mathbf{x}_k \mathbf{x}_k^*$ the orthogonal projector corresponding to λ_k .

The outer product of two column vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as in Equation(4):

$$\mathbf{x} \mathbf{y}^* = \begin{pmatrix} x_1 \bar{y}_1 & \dots & x_1 \bar{y}_n \\ \dots & \dots & \dots \\ x_n \bar{y}_1 & \dots & x_n \bar{y}_n \end{pmatrix} \quad (4)$$

Since the eigenvalues are real they can be sorted $\lambda_1 \geq \dots \geq \lambda_n$ and thus can help to identify the dominant substructures in a network, since the eigenvalues are the weights of the eigenspaces.

Following the concept of eigenvector centrality as described by Wasserman and Faust [10] we identify the most central vertex v_m in a graph $G(V, E)$ by its absolute value $|x_{1,m}|$ of the eigenvector component corresponding to the largest eigenvalue λ_1 . This also holds for the most central vertices in each substructure identified by the

eigenvectors. These central vertices play a major role when explaining the clustering algorithm we propose.

3 The Complex Hermitian Adjacency Matrix

Taken a network as digraph $G(V, E, w)$ without self-references and vertices $v_k \in V$ connected by edges $e_k \in E$ which have weights w , in the representation of a weighted adjacency matrix $A \in \mathbb{R}$, this matrix is transformed into the Hermitian matrix H by Equation(5)

$$H = (A + iA^T)e^{-i\frac{\pi}{4}}. \quad (5)$$

Calculating the eigensystem of this matrix we get a set of non-zero eigenvalues Λ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ and their corresponding complex eigenvectors $\mathbf{X}$ with $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. For a complex Hermitian matrix H with $h_{kk} = 0 \quad \forall k$ (no self-reference) at least one eigenvalue has to be positive due to the fact that $tr(H) = \sum_{k=1}^n h_{kk} = \sum_{k=1}^n \lambda_k = 0$ and some negative.

4 Clustering Within the Eigensystem

The goal of spectral clustering is to find clusters and assign each vertex to a cluster. Following and adapting the k-means approach we assign a vertex to a cluster center node if the distance between these two in the derived vector space is minimal. We use vertices as cluster centers and do not calculate theoretical cluster centers after initialization.

Using Equation(3) and given the eigensystem of H we take the positive eigenvalues $\lambda_1^+, \lambda_2^+, \dots, \lambda_l^+$ and their corresponding normalized eigenvectors $\mathbf{x}_1^+, \mathbf{x}_2^+, \dots, \mathbf{x}_l^+$ and build the row matrix $R_{n \times l}$

$$R_{n \times l} = \begin{pmatrix} \lambda_1^+ \mathbf{x}_{11}^+, \dots, \lambda_l^+ \mathbf{x}_{1l}^+ \\ \vdots \\ \lambda_1^+ \mathbf{x}_{1n}^+, \dots, \lambda_l^+ \mathbf{x}_{ln}^+ \end{pmatrix} \quad (6)$$

The eigenvectors are weighted by their eigenvalues to make the relevant structures more prominent. We take the positive eigenvalues, because these best represent structure building processes within a network with no self-reference, and the corresponding eigenvectors. Simulations have shown this (see [3]). From this matrix representing an l -dimensional subspace of the original vector space of H , we build the symmetric matrix $S_{n \times n} = RR^*$ as the inner product matrix. The central vertices as identified by the maximum absolute value within each eigenvector (see Section 2) strongly influence this calculation. The diagonal elements of S give the norm of each row of R . Thus relevant vertices will show a high diagonal value. Each row of R represents each of the n vertices in the l -dimensional subspace. As seen above the minimal distance between any two vertices can be seen in the maximum of the real part of the inner product between the two. We thus only need to maximize $2Re(S_{n \times n})$ and assign the vertices of the network to the cluster with cluster center $\mathbf{x}_{j,p_c}^+$ such that a vertex i belongs to a cluster center node p_c if $Re(S_{i,p_c}) = \max_j Re(S_{i,j})$. If a node is not assigned to a cluster center node then it is assigned to itself.

As at least one of the eigenvalues is greater than 0, c and therefore the minimum number of clusters is at least one. The number of c is automatically and reproducibly derived from a search over all rows of S . Since the eigensystem does not change while the matrix is not strongly perturbed, the basis for the clustering remains stable, and thus the cluster assignments remain the same. The only source for error when assigning vertices to clusters lies in the numerical precision chosen. Since computational complexity in this case is dominated by the eigensystem calculation it should be bounded by $O(n^3)$. The clustering itself should be bounded by $O(n^2)$.

5 Application

To show how the method works we have taken a matrix of 39 vertices with 4 non-overlapping clusters. We have some results for larger graphs (> 1000 vertices), but to show the effectiveness of the method we restricted the example to a visually understandable matrix.

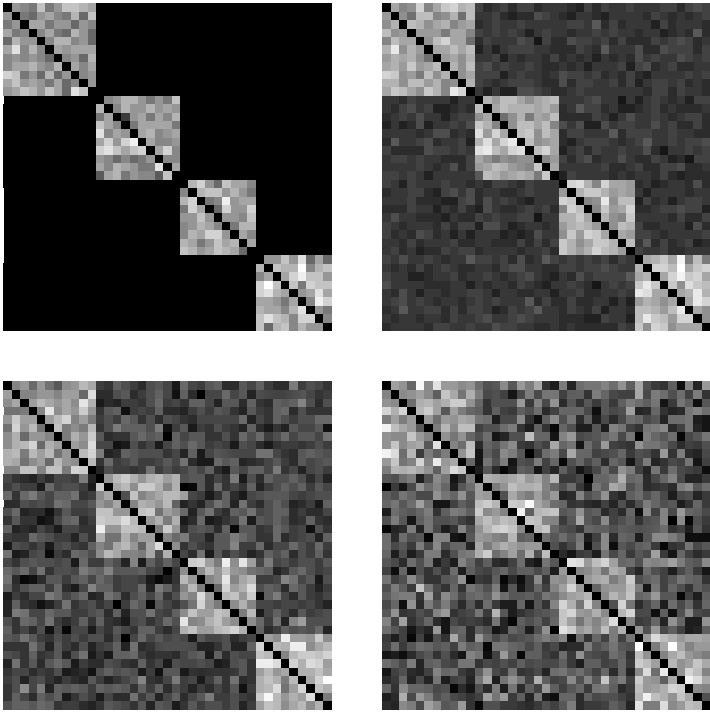


Fig. 1. Matrices with four levels of perturbation

Using this kind of simulation the correct assignment is known. We then have perturbed the matrix in three steps. A normal distribution of perturbation was chosen. For the result see Figure(1). The clusters were generated as four complete graphs of sizes 11, 10, 9 and 9 nodes with a normal distributed in and out degree with mean of 10, 10, 11, and 12 and standard deviation of 2. The perturbations had a normal distribution with mean and standard deviation of (0,0), (5,1), (8,3) and (11,5) from left to right. The upper left matrix is the unperturbed matrix, while the lower right matrix has a perturbation almost of the order of the original data. The eigensystem has been calculated with Mathematica [11].

The result of the proposed method is given in Figure(2). The raised perturbation is again as described above from top left to bottom right. As can be seen when

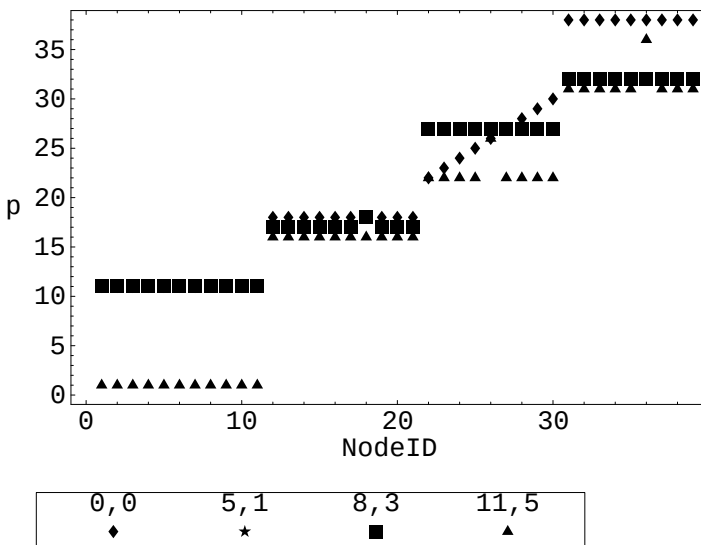


Fig. 2. Quality of clustering

comparing Figures (1) and (2) the assignment never breaks the cluster borders. Even in the case of vertices 23...27 in the upper left hand diagram. Here is a case of numerical precision as these nodes have eigenvector components very close to zero and thus have a zero entry in S . They define isolate clusters. In the top right hand diagram only one node has not been assigned to the cluster center, but has not been assigned to any other cluster also. In the bottom left diagram four nodes have been assigned differently, but within the cluster borders. The bottom right diagram shows only two nodes that have not been assigned. If the cluster center shifts due to perturbation, the assignment is still stable, even if the perturbation is high.

6 Conclusion

In this paper we have presented a method to find the number of clusters automatically from the eigensystem of a complex Hermitian adjacency matrix. It is an improvement in spectral clustering in that no predefinition of the number of clusters and no initialization of cluster centers has to take place. In addition the results are reproducible, thus no multiple runs have to be performed to make sure a good clustering has been achieved.

Some results with overlapping clusters have already been achieved, but a thorough validation is still work in progress. What still needs to be done is to find a mathematical proof, why in the case of an adjacency matrices with zero-trace the positive subspaces yield far better results than the negative. Further evaluation of the method and a comparison with competing methods (e.g. [9]) has to be done.

References

1. C. J. Alpert and S. Yao. Spectral partitioning: The more eigenvectors, the better. In *Proc. of the 32nd ACM/IEEE Conf. on Design Automation*, pages 195–200. ACM Press, 1995.
2. A. Geyer-Schulz B. Hoser. Eigenspectralanalysis of Hermitian Adjacency Matrices for the Analysis of Group Substructures. *Journal of Mathematical Sociology*, 29(4):265–294, 2005.
3. T. Bierhance. Partitionierung von gerichteten Graphen durch Eigensystemanalyse komplexwertiger, Hermitescher Adjazenzmatrizen. Master’s thesis, Universität Karlsruhe (TH), Jun 2005.
4. P. K. Chan, M. D. F. Schlag, and J. Y. Zien. Spectral k-way ratio-cut partitioning and clustering. In *Proc. of the 30th Int. Conf. on Design Automation*, pages 749–754. ACM Press, 1993.
5. T. Choe and C. Park. A k-way Graph Partitioning Algorithm Based on Clustering by Eigenvector. In Marian Bubak, Geert Dick van Albada, Peter M. A. Sloot, and et al., editors, *ICCS 2004: 4th Int. Conf., Kraków, Poland, June 6-9, 2004, Proceedings, Part II*, volume 3037 of *Lecture Notes in Computer Science*, pages 598–601. Springer-Verlag Heidelberg, 2004.
6. M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98), 1973.
7. Martine Schlag Jason Y. Zien, Pak K. Chan. Hybrid spectral/iterative partitioning. In *Proceedings of the 1997 IEEE/ACM international conference on Computer-aided design*, pages 436 – 440. IEEE Computer Society, 1997.
8. R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.
9. A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. In S. Becker T. G. Dietterich and Z. Ghahramani, editors, *Advances in Neural Inform. Proc. Systems*, volume 14, pages 849–856. MIT Press, 2002.
10. Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge, 1 edition, 1999.
11. Inc. Wolfram Research. *Mathematica*. Wolfram Research, Inc., version 5.1 edition.

Online Dial-A-Ride Problem with Time Windows: An Exact Algorithm Using Status Vectors

Anke Fabri and Peter Recht

Universität Dortmund, Germany

a.fabri@wiso.uni-dortmund.de

p.recht@wiso.uni-dortmund.de

Summary. The Dial-A-Ride Problem (DARP) has often been used to organize transport of elderly and handicapped people, assuming that these people can book their transport in advance. But the DARP can also be used to organize usual passenger or goods transportation in real online scenarios with time window constraints. This paper presents an efficient exact algorithm with significantly reduced calculation times.

1 Introduction

The problem consists in routing a vehicle that serves a set N of demands growing in time. The road network is represented by a weighted graph $G = (V, E, d)$ whose vertices $v \in V$ correspond to the possible pickup and delivery points and whose arcs $e \in E$ represent shortest paths between these points. The cost of each arc is given by a function $d: E \rightarrow \mathbb{R}^+$, indicating the length of the shortest path. The planning horizon is denoted by the interval $[0, T]$. Demands for transportation pop up dynamically, without any information in advance about time of appearance, location or quantity of incoming demands. Each demand $i \in N$ consists of a pickup point $u_i \in V$ with a pickup time window $[a_i, b_i] \subset [0, T]$, a delivery point $v_i \in V$ with a delivery time window $[c_i, e_i] \subset [0, T]$ and the quantity of goods to be transported. The vehicle is characterized by its starting point s where it is located at time 0, its end point z where it should be at time T and its capacity k . When at any time a new demand pops up, it must be decided immediately whether the demand is accepted or rejected. If it is accepted, a valid tour has to be found that serves all accepted demands meeting all time windows exactly. The vehicle may wait for service. The objective consists of two parts: first, as many demands as possible shall be served. Second, the traveled distance needed to serve all accepted demands shall be minimized.

This paper presents a special graph structure to handle the DARP and sketches the solution algorithm. The results show that calculation efforts and therefore calculation times are reduced significantly.

2 The Status Vector Graph

The basic status vector graph $G'_n = (V'_n, E'_n, d')$ forms the basis for a dynamic programming solution sequencing the pickup and delivery points of all n accepted demands. Each accepted demand $i \in N$ is characterized by its status $j(i) \in \{0, 1, 2\}$. Here, $j(i) = 0$ means ‘demand i is still waiting for service’, $j(i) = 1$ denotes ‘demand i has been picked up, but not yet delivered’ and $j(i) = 2$ says ‘demand i has been served completely’. A status vector is a vector $(j(i_1), j(i_2), \dots, j(i_n))$ that represents the status of each demand $i \in N$ with $\{i_1, i_2, \dots, i_n\} = N$. The set of all possible status vectors of all demands $i \in N$ forms the set of vertices V'_n . Two vertices $v', v'' \in V'_n$ are connected by an edge $(v', v'') \in E'_n$ whenever vector v'' can be obtained from vector v' by adding 1 to exactly one vector element of v' . This change corresponds to a pickup or delivery of one demand, implying a vehicle movement between two nodes in G . The arc (v', v'') is therefore weighted with the corresponding costs (see [1]). Note that the single status vector does not contain any information about the vehicle’s location. The source vertex is the status vector representing the actual status when the calculation is started, the sink vertices are $(2, \dots, 2)$. Figure 1

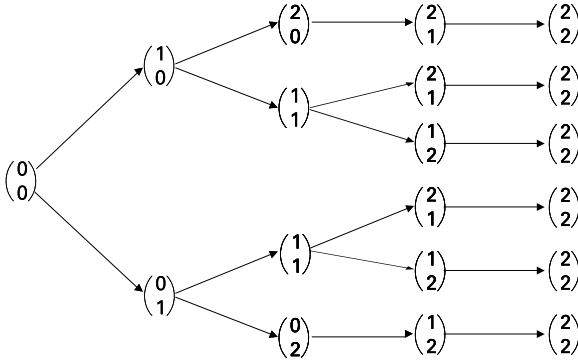


Fig. 1. Basic Status Vector Graph for two Demands

shows G'_2 for both demands still waiting for pickup. Calculating a tour now consists in finding a shortest path from the source to one of the sink vertices subject to time window and capacity constraints.

Attention is attracted by the fact that many status vectors are identical. Calculation times could be significantly reduced, if these multiple status vectors could be merged to obtain a graph of reduced size, say $redG'_n = (redV'_n, redE'_n, red d')$. Due to the cost function depending on the vehicles location, two identical vectors can only be merged if also the vehicle location is identical. Hence, one more piece of information has to be stored. See figure 2 for an example of $redG'_2$ with pickup places A, B and delivery places C, D for demand 1, 2, both still waiting for service.

Regarding the number of nodes of $redG'_n$, one must state that it is worth storing the location additionally:

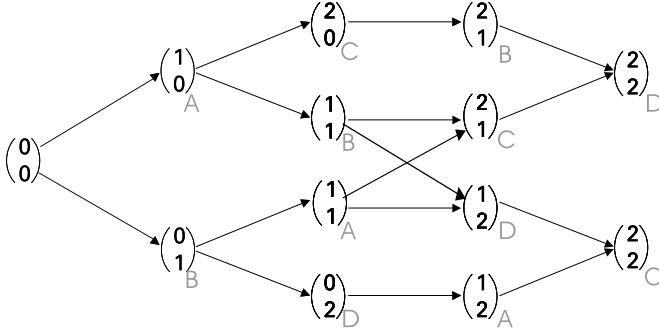


Fig. 2. Graph ${}_{red}G'_2$ for both demands still waiting for pickup

Lemma 1. For a set of n demands not yet served the following equations hold:

$$|V'_n| = \sum_{j=0}^n \frac{n!}{(n-j)!} \left(1 + \sum_{t=1}^j \frac{(j+t)!}{2^t(j-t)!t!} \right)$$

$$|{}_{red}V'_n| = 1 + \sum_{k=1}^n \sum_{l=0}^{k/2} \frac{n! (k-l)}{l! (k-2l)! (n-k+l)!} + \sum_{k=n+1}^{2n} \sum_{l=k-n}^{k/2} \frac{n! (k-l)}{l! (k-2l)! (n-k+l)!}$$

The proof is technical and can be found in [3]. It essentially uses the special structure of ${}_{red}G'_n$ which allows a recursive calculation of the number of nodes.

Table 1 shows the number of nodes of both graphs for $n = 1..6$ demands. It is

Table 1. Number of nodes in the basic / reduced status vector graph for up to 6 demands

n	1	2	3	4	5	6
$ V'_n $	3	19	271	7365	326011	21295783
$ {}_{red}V'_n $	3	13	55	217	811	2917
reduction	0%	31,58%	79,70%	97,05%	99,75%	99,99%

obvious that the reduced graph may lead to significantly reduced calculation times considering the savings in the number of nodes. The following section treats the algorithm highlighting the implementation to realize this reduction.

3 The Algorithm

To find a shortest path from the source to one of the sink vertices in ${}_{red}G'_n$ satisfying the time window and capacity constraints, the A*-algorithm is used. The A*-algorithm bases on the Dijkstra algorithm but expands the cost function by adding an estimated value for the remaining distance to the sink. The optimal solution is obtained whenever the distances satisfy the triangle inequality and the estimated distance to the sink does not overestimate the real value (see [5]). With a given estimate that calculates the maximum distance of all left demands assuming that only one of them was left to be served (see [2]) and the distances between two vertices consisting of the shortest paths, these assumptions are met. Due to the special structure of ${}_{red}G'_n$, the graph does not have to be given in advance but is constructed iteratively and only 'as far as necessary' during the algorithm. As shown in [2], this algorithm performs well on G'_n , especially with the presented additional feasibility checks.

To adapt the algorithm to ${}_{red}G'_n$, it is necessary to check quickly whether the status vector recently generated is really new or just a copy of another one. Therefore, an efficient search structure is crucial.

3.1 Efficient Search Structure

During the A*-algorithm, the generated nodes of ${}_{red}G'_n$ have to be searched for two criteria: the A*-algorithm needs to find the node with minimal (partly estimated) cost that can be expanded next, while the graph structure itself leads to searching for a node copy. Therefore, all nodes are concatenated twice: in the order of their costs and in the order of their status vectors. To give a unique numerical order to the status vectors, each of them is assigned its status number tuple, consisting of the actual place and the representation of the status vector in the decimal system, interpreting it as a number in the number system modulo 3. This leads to efficiently comparable nodes.

For a quick search in both node chains a double skip list is constructed, as skip lists are nearly as efficient in search as a balanced tree but significantly better when inserting and deleting elements (see [6]). A skip list is an ordered linked list with randomly added additional links. Every list element gets a random number of link levels with a minimum of one and a maximum of 6 levels and growing probability for a lower number of levels. Every element is linked on each of its levels to the next element having that level. The source vertex has the maximum number of level. To search this ordered list for a special value, one starts at the source and follows the highest level links until the value was exceeded with the next step. Then the search is continued one level below and so on. To be able to search for both cost and status number, each level has two links: one to the node with the next best cost having the same level and one to the node with the next higher status number having the same level. Figure 3 shows a partly generated ${}_{red}G'_2$ with status numbers and locations and its representation as a skip list. For a clearer intuition, only the status links are shown.

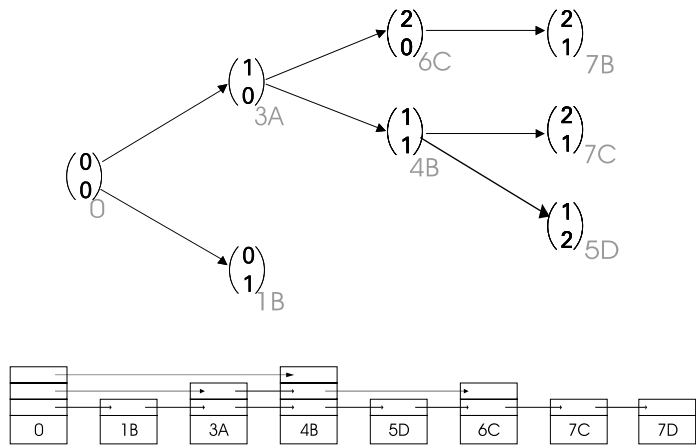


Fig. 3. Example of a status skip list for a graph under construction with 2 demands

4 Results and Conclusion

To test the algorithm, the further modified Li and Lim test instances for dynamic pickup and delivery vehicle routing with time windows with about 100 places to visit are used (see [4]). These test instances are changed in the way that only one vehicle is used. The results are given as an average value of the "lc", "lr" and "lrc" test cases representing concentrated, random or a mixture of concentrated and random located places to visit. Due to the capacity constraints, only the very first demands can be served. The algorithm running on the basic status vector graph is stopped as soon as 10 million nodes have been generated. Table 2 shows the average number of orders treated completely until then, the maximal and average calculation times to accept or reject one order and the number of nodes generated to treat these orders, divided in generated nodes over all, unfeasible nodes, multiple nodes (reduced to one) and expanded nodes. The algorithm running on the reduced status vector graph is stopped at the same number of orders to compare number of nodes and calculation times in seconds.

Table 2 shows that the number of nodes and the calculation times could be significantly reduced. This significant reduction is due to the fact that many nodes have been recognized as multiple nodes and could be merged into only one node implying that only this one node is further expanded. To conclude, we can summarize that it is really worth relying on the more complicated data structure of the reduced graph due to the impressing reduction of calculation times.

Table 2. Nr of nodes and calculation times for the test cases "100"

	nr of orders	graph	order time		nr of nodes			
			average	max.	over all	unfeasible	multiple	expanded
lc	26,82	basic	18,26	2159,12	345835,88	325812,41		165198,76
		reduced	0,06	7,93	17278,53	56712,94	5105,65	14378,53
lr	31,13	basic	39,05	5806,79	428855,43	86143,39		140869,30
		reduced	0,01	0,88	6675,78	7728,61	6343,13	5123,83
lrc	33,50	basic	29,64	6108,28	316402,88	161555,25		120615,19
		reduced	0,01	0,84	9735,31	19844,81	6951,31	7575,31

References

1. Caramia, M., Italiano, G. F. , Oriolo, G., Pacifici, A. and Perugia, A. (2001) Routing a fleet of vehicles for dynamic combined pickup and delivery services. *Proceedings of the Symposium on Operation Research 2001*, pp 3-8. Berlin/Heidelberg: Springer-Verlag

2. Fabri, A. , Recht, P. (2006) On dynamic pickup and delivery vehicle routing with several time windows and waiting times. *Transportation Research Part B*, 40, pp335-350. Elsevier

3. Fabri, A. (2006) Die Größe der Statusgraphen des DARP unter Ladebedingungen. *Working Paper. Universität Dortmund, Fachgebiet Operations Research*

4. Fabri, A. (2005) Benchmarkprobleme, <http://www.wiso.uni-dortmund.de/orwi/de/content/team/mitarbeiter/Benchmarks.html>, 19.07.2006

5. P. E. Hart, N. J. Nilsson, B. Raphael: A Formal Basis for the Heuristic Determination of Minimum Cost Paths; IEEE transactions of systems science and cybernetics, Vol. 4, No.2, 1968

6. Pugh, William (1990) Skip lists: a probabilistic alternative to balanced trees. *Commun. ACM, vol. 33 nr. 6* ACM Press,pp. 668–676. New York: ACM Press

Operational and Credit Risk

Die Anwendung des Verlustverteilungsansatzes zur Quantifizierung operationeller Risiken

Frank Beekmann and Peter Stemper

WestLB, Düsseldorf, Germany

Frank.Beekmann@westlb.de

Dr.Peter.Stemper@westlb.de

Summary. Der vorliegende Beitrag zeigt die in der WestLB angestrebte Anwendung des Verlustverteilungsansatzes, der zur Quantifizierung operationeller Risiken nach der Basel II-Rahmenvereinbarung herangezogen werden kann. Der Verlustverteilungsansatz gehört zur Klasse der fortgeschrittenen Messverfahren für operationelle Risiken und stellt in dieser den in der Bankenlandschaft am weitesten verbreiteten Ansatz dar.

Kern des Ansatzes ist die Bestimmung der Verteilung einer Summenzufallsvariablen, deren Summanden Zufallsvariablen sind, mit denen Schäden aus dem operationellen Betrieb einer Bank beschrieben werden. Die Anzahl der Summanden ist ebenfalls eine Zufallsvariable, welche die Anzahl aufgetretener Schäden darstellt. Aus dieser zusammengesetzten Verteilung lassen sich dann operationelle Risiken etwa als 0.999-Quantil quantifizieren.

Im Beitrag wird erläutert, wie die Einzelschadenverteilungen vor dem Problemfeld geringer Datendichte bestimmt werden, wie die Daten aus Szenarioanalysen integriert werden, wie Summenvariablen unterschiedlicher Risikokategorien aggregiert werden sowie die Bestimmung der Summenverteilung unter Einbezug von Versicherungsleistungen mittels einer Monte Carlo Simulation.

1 Einleitung

Die Bankenbranche ist derzeit mit der Umsetzung von regulatorischen Anforderungen beschäftigt, die von dem international zusammengesetzten "Baseler Ausschuss für Bankenaufsicht" in dessen überarbeiteter Rahmenvereinbarung (Basel II) dokumentiert worden sind. Darin nehmen Anforderungen zur Quantifizierung von Risiken im Bankgeschäft breiten Raum ein, wobei den operationellen Risiken eine bedeutende Rolle zugesprochen wird. Diese Risiken sind definiert als "die Gefahr von Verlusten, die ihre Ursache in der Unangemessenheit oder dem Versagen von Geschäftsprozessen, Technologie oder des Personals einer Bank haben oder als Folge von externen Ereignissen auftreten" [1].

Das Risiko kann beispielsweise durch einen operationellen Value at Risk (OpVaR) gemessen werden, welcher ein Gesamtschadenwert ist, der mit einer vorgegebe-

nen Wahrscheinlichkeit nicht überschritten wird. Der 0,999-OpVaR stellt das zur Deckung von Schäden aus operationellem Risiko notwendige regulatorische Kapital dar. Der 0.9995-OpVaR ist dagegen größer und wird zur internen Steuerung der Bank verwendet (ökonomisches Kapital).

Um diesen OpVaR zu bestimmen, gibt es mehrere unterschiedliche Ansätze. Die einfachsten Ansätze empfehlen, den OpVaR als einen vom Geschäftsfeld der Bank abhängigen Prozentsatz des operativen Bruttoertrages zu berechnen. Alternativ kann ein sogenannter Advanced Measurement Approach (AMA) angewendet werden, der nach Basel II zwar methodisch nicht näher erläutert wird, jedoch qualitative Anforderungen erfüllen muss. Daraus haben sich einige Klassen von Ansätzen entwickelt, von denen der Verlustverteilungsansatz der am weitesten verbreitete ist. Wesentliche Anforderungen für ein Quantifizierungsmodell sind die Nutzung interner und externer Schadendaten, Szenariodaten sowie Korrelationsannahmen.

Im Folgenden wird die Ausgestaltung dieses Ansatzes bei der WestLB erläutert.

2 Grundmodell

Das Modell zur Bestimmung des OpVaRs ist vor dem Hintergrund der Schadensmatrix zu sehen. Die auftretenden Schäden in der WestLB werden der Baseler Richtlinie folgend nach Geschäftsfeldern (Business Lines) und Schadenereignistypen (Event Types) kategorisiert. Die Geschäftsfelder, in denen die WestLB überwiegend tätig ist, und die Ereignistypen, in denen Schäden aufgetreten sind, sind in Tabelle 1 dargestellt.

Table 1. Schadensmatrix

	Business Disrup- tion & System Fail- ures	Execu- tion, Delivery & Process Manage- ment	Clients, Prod- ucts & Business Practices	Employ- ment Practices & Work- place safety	Dam- age to phys- ical assets	Inter- nal Fraud	Exter- nal Fraud
Asset Management							
Commercial Banking							
Corporate Finance							
Payment & Settlement							
Trading & Sales							

Für jede Zelle dieser Schadensmatrix wird ein eigener OpVaR berechnet, sofern hinreichend viele Schadendaten in den jeweiligen Zellen vorhanden sind. Wo dies nicht der Fall ist, werden auch Daten verschiedener Zellen vereinigt, wenn dies etwa wie im Fall Internal und External Fraud auch aus der Anwendungssicht vertretbar ist. Die OpVaR-Bestimmung orientiert sich an dem klassischen kollektiven Modell der Versicherungswirtschaft [2]. Hierbei wird eine Gesamtschadenvariable L durch eine Zufallsvariable N, welche die Anzahl der Schäden, und unabhängig und identisch

verteilten Zufallsvariablen $X_i (i = 1, \dots, N)$ zur Beschreibung der Schadenhöhen wie folgt gebildet

$$L = \sum_{i=1}^N X_i \quad (1)$$

Das Ziel ist nun, eine Verteilung der Gesamtschadenvariablen zu bestimmen, um daraus die Kenngrößen Erwartungswert sowie das 0,999- und das 0,9995-Quantil abzuleiten. Der Erwartungswert gibt dann den Expected Loss in dieser Zelle an, die Quantile ergeben jeweils einen OpVaR mit entsprechender Wahrscheinlichkeit.

Die einzelnen OpVaRs der Zellen müssen schließlich zu einem Gesamt-OpVaR aggregiert werden. Dies geschieht im einfachsten Fall durch Addition, was allerdings die sehr konservative Annahme widerspiegelt, dass in allen Zellen gleichzeitig der jeweils hohe Quantilwert als Gesamtschaden eintritt.

3 Bestimmung der Einzelverteilungen

Das erste Problem ist die Bestimmung der Verteilungen für die Schadenanzahlverteilung sowie für die Schadenhöheverteilung.

Die Schadenanzahl N wird durch eine Poissonverteilung modelliert, der zugehörige Parameter λ wird durch die mittlere Schadenhäufigkeit der internen Schadendaten geschätzt. Dabei ist zu beachten, dass N nur die Anzahl der Schäden oberhalb der Erfassungsgrenze interner Daten beschreibt. Aufgrund der geringen Datendichte kann es außerdem vorkommen, dass nicht in allen Zellen interne Daten zur Schätzung vorliegen. In diesem Fall wird konservativ angenommen, dass mindestens ein Schaden aufgetreten ist.

Die Bestimmung der Schadenhöheverteilung umfasst zwei Problemfelder. Erstens die Auswahl eines geeigneten Verteilungstyps, zweitens die Schätzung der zugehörigen Parameter aus den Daten.

Als Verteilungstypen werden einfache Verteilungstypen wie die Lognormal- oder Weibullverteilung verwendet sowie gesplittete Verteilungen, die sich an einer Splittgrenze aus einem Body- und einem gewichteten Tailbereich zusammensetzen. Während die Bodyfunktion als lognormalverteilt angenommen wird, finden für die Tailverteilungen die Lognormal-, Weibull- oder Gammaverteilung sowie die aus der Extremwerttheorie stammende Generalised Pareto Distribution (GPD) Anwendung [3]. Die Auswahl des richtigen Typen geschieht durch Vergleich von P-Values von statistischen Tests wie dem Kolmogorov-Smirnov-Test, dem Anderson-Darling-Test und dem Cramer-von-Mises-Test. Darüber hinaus werden graphische Goodness-of-Fit-Methoden wie der QQ-Plot oder der Vergleich der empirischen und der analytischen Verteilungsfunktion angewendet.

Für die Parameterschätzung wird der Maximum-Likelihood-Schätzer verwendet. Die Daten für die Schätzung stammen aus externen Quellen wie dem ORX-Datenkonsortium, das ein Zusammenschluss von derzeit 27 internationalen Banken ist, welche sich ihre anonymisierten Schadendaten ab einer Grenze von 20.000 Euro gegenseitig zur Verfügung stellen. Durch die Erfassungsgrenze ist bei der Schätzung die trunkierte Verteilungsfunktion zu verwenden, d. h. als zu maximierende Loglikelihoodfunktion muss

$$l = \sum_{i=1}^R \ln \frac{f(x_i)}{1 - F(20.000)} \quad (2)$$

gewählt werden, wobei $x_i (i = 1, \dots, R)$ die Schadenhöhen, R die Anzahl der Datenpunkte, f die Dichte der Verteilung und F die zugehörige Verteilungsfunktion bezeichnet.

Zusätzlich werden auch Szenariodaten für die Schätzung verwendet. Diese entstammen Szenarioanalysen, die innerhalb der Bank durchgeführt werden. Dabei legen Experten mögliche Schadenereignisse fest und schätzen die Höhe des Schadens sowie eine zugehörige Eintrittswahrscheinlichkeit. Die Szenariopunkte werden dann als zusätzliche Werte beim Maximum-Likelihood-Schätzer berücksichtigt, allerdings mit einer Gewichtung, die sich aus der angegebenen Eintrittswahrscheinlichkeit ergibt.

4 Bestimmung des OpVaRs

Nach erfolgter Parametrisierung findet die Bestimmung der OpVaR-Werte durch eine Monte Carlo Simulation statt. Hierbei wird eine hinreichend große Anzahl von Szenarien (mindestens 1 Mio) simuliert, die jeweils die Gesamt-schadenssumme eines Jahres angeben. Für jedes Szenario wird zunächst eine Schadenanzahl n als Instanz der Zufallszahl N erzeugt, anschließend werden n Zufallszahlen für die aufgetretenen Schäden generiert. Hierbei ist zu beachten, dass die Zufallszahl ab der internen Erfassungsgrenze gezogen werden. Die Summe dieser Einzelschäden bedeutet dann gerade die Simulation einer Gesamtschadenssumme. Aus diesen so simulierten Szenarien ergibt sich eine Verteilung von L , aus der sich die Kennzahlen Erwartungswert und Quantile als OpVaRs ermitteln lassen.

Bei der Ziehung der Zufallsvariablen für die Schadenhöhe werden Versicherungsleistungen berücksichtigt. Hierzu wird für jede simulierte Schadenhöhe eine auf $[0,1]$ gleichverteilte Zufallszahl generiert, welche durch Vergleich mit der vorgegebenen Versicherungseintrittswahrscheinlichkeit das Ereignis "Versicherung greift" anzeigt. Tritt dieses Ereignis ein, so wird die Schadenhöhe um die Versicherungsleistung etwa als Überschuss über den Selbstbehalt korrigiert.

Werden keine Versicherungsleistungen berücksichtigt, kann der OpVaR auch approximativ analytisch bestimmt werden durch

$$OpVaR(p) = F^{-1} \left(1 - \frac{1-p}{E[N]} \right) \quad (3)$$

wobei p das Sicherheitsniveau (0.999 bzw. 0.9995) ausdrückt und F^{-1} die Umkehrfunktion der Verteilungsfunktion der Schadenhöhevariable. Diese Formel basiert auf einer Eigenschaft von sub-exponentiellen Verteilungen, wie die oben angegebenen [4]. Für $E[N]$ kann der beobachtete Mittelwert der Schadenanzahl der internen WestLB-Daten gewählt werden. Hierbei ist allerdings zu beachten, dass sich dieser Mittelwert nur auf die Schäden ab der Schadenerfassungsgrenze bezieht und somit für F eine bedingte Verteilung verwendet werden muss.

Auch kann der Erwartungswert selbst unter Berücksichtigung von Versicherungsleistungen analytisch bestimmt werden. Dazu wird das Ergebnis aus der Versicherungsmathematik herangezogen, dass für den Erwartungswert von L gilt:

$$E[L] = E[N]E[X] \quad (4)$$

wobei X die Zufallsvariable der Schadenhöhe mit Versicherungsleistung bezeichnet.

Das oben beschriebene Verfahren wird in dieser Form zur Bestimmung der Einzelzellen verwendet. Um einen aggregierten OpVaR zu erhalten, muss das Verfahren erweitert werden. In der WestLB werden aggregierte OpVaRs nur bezüglich der Business Lines (BL) modelliert. Die Summe dieser BL-OpVaRs ist der Gesamt-OpVaR der WestLB. Für die Bestimmung des BL-OpVaRs werden in jedem Szenario der Monte-Carlo-Simulation die ermittelten Schadensummen in den zugehörigen Ereignisklassen zu einer BL-Schadensumme addiert. Damit ergibt sich eine Verteilung des Verlustes in einer Business Line, woraus die Kennzahlen analog zum Einzelzellenfall ermittelt werden. Dieses Vorgehen basiert implizit auf die Annahme, dass die Ereignisklassen unabhängig voneinander sind. Um hierbei Korrelationsannahmen zwischen den Schadenereignisklassen einfließen zu lassen, werden die Zufallszahlen für die Schadenanzahlen als Vektor mit korrelierten Komponenten gezogen. Dazu wird eine Gaußsche Copula verwendet [5].

5 Weiterverarbeitung der OpVaRs

Der oben beschriebene Ansatz kann genutzt werden, um das operationelle Risiko zu quantifizieren. Die derart ermittelten Business-Line-OpVaRs werden dann auf die zu den Business Lines gehörenden Geschäftsbereiche allokiert.

Anschließend erfolgt eine bereichsspezifische Adjustierung des OpVaRs auf Basis einer Kennzahl für das Management operationeller Risiken.

Der nach dieser Adjustierung ermittelte Wert für das ökonomische Kapital geht in die Steuerungssysteme der Bank ein und dient damit neben der Messung des Risikos zur risikoadäquaten Steuerung der operativen Einheiten.

References

1. Basel Committee on Banking Supervision (2004) International convergence of capital measurement and capital standards, a revised framework
2. Klugman SA, Panjer HH, Willmot GE (2004) Loss models. From data to decision. Wiley, Holboken
3. Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events. Springer, Berlin Heidelberg
4. Böcker K, Klüppelberg C (2005) Operational VAR: a closed-form approximation, RISK Magazine, December, 90-93
5. Frachot A, Roncalli T, Salomon E (2005) Correlation and diversification effects in operational risk modelling. In: Davis E (ed) Operational risk. Practical approaches to implementation. Risk Books, London

Production and Supply Chain Management

Betriebskennlinien-Management als Performancemessungs- und -planungskonzept bei komplexen Produktionsprozessen

Dirk Eichhorn¹ and Alexander Schömig²

¹ Qimonda AG, München, Germany

Dirk.Eichhorn@qimonda.com

² Infineon Technologies AG, München, Germany

Alexander.Schoemig@infineon.com

1 Einleitung

Schwerpunkt des traditionellen Produktions-Controlling ist der Bereich der Kosten. In Abhängigkeit davon, welche Rolle Lieferfähigkeit und Liefertreue in der Branche spielen, werden output-bezogene Performance-Indikatoren eingeführt. Während damit im Allgemeinen nur Symptome diagnostiziert werden können, stellt sich die Frage nach den Gründen von Planabweichungen und Möglichkeiten zur Verbesserung der Lieferperformance. Um Produkte gemäß Kundenwunsch auszuliefern sind mit Sicherheit die Durchlaufzeit und deren Varianz von Interesse. Eine ex-post Messung dieser Größen ist hilfreich; wichtiger ist jedoch eine Verringerung der Werte. Daher ist es zweckdienlich, die logistische Leistung bereits innerhalb des Fertigungsablaufs zu messen, statt an dessen Ende. Somit sollte es leichter fallen, Abweichungen früher zu erkennen und rechtzeitig Gegenmaßnahmen zu ergreifen. Mit Hilfe geeigneter Daten ist möglich, nicht nur die Durchlaufzeit zu verstehen, sondern auch Durchsatz- und Kostenthemen zu erklären und diese für planerische Zwecke zu verwenden.

Das vorzuschlagende Konzept basiert auf der Warteschlangentheorie: Es wird ein Zusammenhang zwischen der Auslastung des Fertigungssystems und dem sogenannten "Flussfaktor", dem Quotienten aus Durchlaufzeit und theoretischer Minimaldurchlaufzeit aufgestellt. Als wichtiger Einflussfaktor wird die Variabilität "Alpha" des Systems eingeführt. (Die Grundlagen finden sich bei Hopp/Spearman [5] und den Literaturstellen darin, sowie bei Engelhardt [4].) Ausgehend von einem Single- oder Multi-Server-Ansatz ist die Aggregation zur Gesamtlinie mittels weniger Datensätze möglich. Der funktionale Zusammenhang zwischen Durchsatz und Durchlaufzeit einer Fabrik kann zur Visualisierung der Performanceentwicklung über der Zeit verwendet werden. Darüber hinaus kann das System mit Investitions- und Kostenplanungsmodellen verknüpft werden. Auf einer detaillierteren Workcenter-Ebene kann die Wirkung von Maßnahmen auf die Gesamtdurchlaufzeit und die Kostenperformance abgeschätzt werden.

Das beschriebene Konzept wird bei der Infineon AG und Qimonda AG seit längerer Zeit erfolgreich eingesetzt. In dieser Arbeit wird eine Verfeinerung des Systems

beschrieben, die zu einer realistischeren Abbildung des Fertigungssystems basierend auf einzelnen Toolgruppen–Betriebskennlinien beruht.

2 Konzeptentwurf

2.1 Benötigte Daten

Fertigungsprozesse in der Halbleiterindustrie sind von einer hohen Komplexität gekennzeichnet. Daraus resultiert eine lange Durchlaufzeit. Das Prinzip der Werkstattfertigung mit seinen Beständen vor den Workcentern erlaubt ein Weiterlaufen der Fertigung an Maschinen, die sich im Prozessablauf nach einer Anlage, die von Problemen betroffen ist, befinden. Die Messung der Produktionsleistung anhand des Outputs würde daher zu einer sehr späten Reaktion auf Performanceschwankungen in der Linie führen. Aus diesem Grund werden alle durchgeführten Prozessschritte erfasst. Vereinfachend ist es ausreichend, immer wiederkehrende Schritte innerhalb eines Produktdurchlaufes zu zählen und von diesen auf die gesamten Prozessschritte zu schlussfolgern. In der Halbleiterindustrie repräsentiert die Lithographie einen Prozessschritt, der hinreichend häufig und regelmäßig im Arbeitsplan eines Produktes auftritt. Anhand der Gesamtzahl geleisteter Litho–Schritte geteilt durch die gemäß Arbeitsplan zu erwartende Zahl wird eine so genannte Daily Going Rate (*DGR*) berechnet. Es handelt sich dabei um ein Maß für den Durchsatz, konkret eine theoretisch gefertigte Anzahl an Produkten, unabhängig davon, ob diese aus dem Fertigungssystem ausgeschleust wurden.

Neben dieser Durchsatzgröße ist es nahe liegend, den Bestand (*WIP*) im System zu erfassen. Durch die Verknüpfung beider Werte mittels Little’s Law kann die dynamische Durchlaufzeit (*dynCT*) bestimmt werden. Im Gegensatz zur output–bezogenen Durchlaufzeit reagiert diese Größe beispielsweise sofort auf erhöhte Einschleusungen in das System und zeigt daher Abweichungen vom Sollzustand früher an.

Um eine von der Fertigungstechnologie unabhängige Beurteilung der Durchlaufzeit zu ermöglichen, empfiehlt sich eine Normierung. Der Bezug auf die Raw Process Time (der theoretischen Mindestprozesszeit je Fertigungsschritt) ergibt den Flussfaktor (*FF*), im angelsächsischen Raum vorzugsweise *X–Factor* genannt. Aus diesem Grund müssen die Raw Process Times der Produkte vorliegen.

$$FF = \frac{dynCT}{RPT} \quad (1)$$

2.2 Warteschlangentheorie

In einer Werkstattfertigung lassen sich die einzelnen Anlagen als Server und die zu verarbeitenden Teile bzw. Lose als Forderungen in einem Bediensystem bzw. Bedienungnetz interpretieren. In [2] und [3] wird die Ableitung der Betriebskennlinie aus den grundlegenden Formeln der Warteschlangentheorie dargelegt. Dabei wird gezeigt, dass das sehr einfache Ergebnis des Single–Server–Falles $G/G/1$ zur Approximation des Multi–Server–Falles $G/G/s$ verwendet werden kann. Abbildung 1 zeigt beispielhaft die entsprechenden Abweichungen. In einem weiteren Abstraktionsschritt wird angenommen, dass die Single–Server–Formel

$$FF = \alpha \frac{UU}{1 - UU} + 1 \quad (2)$$

auch zur Abbildung des gesamten Bedienungsnetzes, d.h. der Fertigungslinie geeignet ist. Diese Annahme ist qualitativ korrekt. Der prinzipielle Verlauf mit einer Untergrenze bei $FF=1$ und einem progressiven Verlauf durch den gemessenen Arbeitspunkt hin zu sehr hohen Flussfaktoren bei Annäherung an die Grenzauslastung kann durch Simulationen oder vollständig modellierte Bedienungsnetze bestätigt werden. Mithin kann basierend auf der Messung von Durchlaufzeit respektive Flussfaktor FF und Auslastung UU der Fertigungslinie der Parameter Alpha in Gleichung 2 retrograd ermittelt werden.

In der Praxis zeigt sich jedoch, dass die quantitativen Aussagen dieses Modells unter Umständen signifikant von der Realität abweichen. Hauptgrund dafür ist, dass reale Linien nicht ausbalanciert sind. D.h. die verschiedenen Toolgruppen weisen nicht identische Kapazitäten auf. Daraus resultiert die Schwierigkeit, eine für die Gesamtlinie repräsentative Kapazitätsgrenze festzulegen. Unmittelbar einleuchtend wäre die Verwendung der Bottleneck-Kapazität. Dadurch ergibt sich rechts vom Arbeitspunkt ein Verlauf, der zumindest eine realitätsnahe Aussage bzgl. der Grenzkapazität des Systems (in Stück) zulässt. Das Modell wird jedoch durch das Bottleneck dominiert, sodass der Kurvenverlauf links vom Arbeitspunkt regelmäßig als nicht sinnvoll erkannt werden muss. Unter Inkaufnahme eines unrealistischen Verlaufs rechts vom Arbeitspunkt hat sich in der Praxis die Verwendung der durchschnittlichen Auslastung aller Anlagen als brauchbarer Kompromiss erwiesen. Dennoch weist dieser Ansatz je nach Kapazitätsprofil der Linie deutliche Schwächen auf. Daher sind Überlegungen erforderlich, bei vertretbarem Aufwand eine realitätsnähere

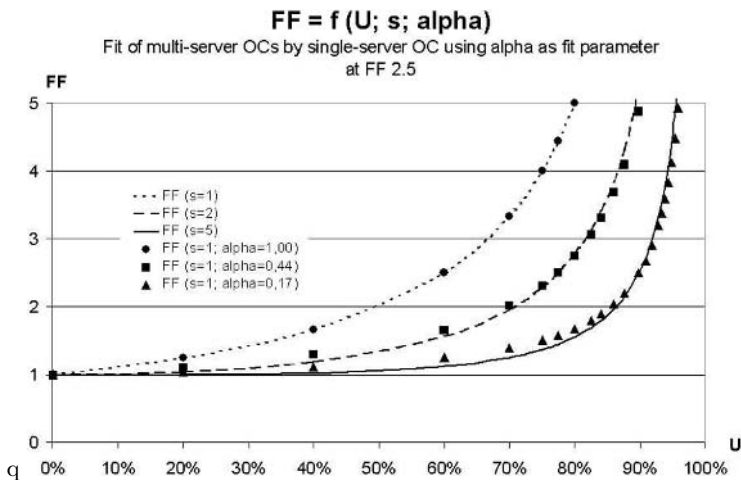


Fig. 1. Single- und Multi-Server-BKL.

Abbildung des Linienverhaltens bei unterschiedlichen Auslastungen zu modellieren. Ausgangspunkt ist dabei weiterhin der Wunsch, den Systemzustand basierend auf Ist-Werten einer bestimmten Periode darzustellen. Der Arbeitspunkt kann also als gegeben betrachtet werden. Für diese ex-post Analyse sind daher aufwändige Bedienungsnetzmodelle mit multidimensionalen Gleichungssystemen nicht erforderlich.

Ideenansatz ist die Gültigkeit von Gleichung 2 für einzelne Maschinen bzw. näherungsweise für Toolgruppen. Alternativ kann für Toolgruppen auch die Multi-Server-Formel

$$FF = \frac{\alpha}{s} \cdot \frac{UU_s^{\sqrt{2(m+1)}-1}}{1 - UU_s} + 1 \quad (3)$$

verwendet werden. Die damit berechneten Betriebskennlinien einzelner Toolgruppen müssen geeignet aggregiert werden, um eine Betriebskennlinie des Gesamtsystems zu erhalten, die alle Teilsysteme abbildet und somit insgesamt realitätsnah ist. Die Überlegung, Einzelkennlinien der Durchlaufzeit als Funktion der stückbasierten Auslastung zu aggregieren ist zunächst naheliegend, erweist sich in der Praxis jedoch aus verschiedenen Gründen (z.B. bei Produktmischen mit unterschiedlichen Arbeitsplänen) als schwierig. Deutlich robuster ist eine Aggregation über die im System befindlichen Bestände. Über Little's Law wird dann die dynamische Durchlaufzeit des Gesamtsystems berechnet:

$$dynCT = \frac{\sum_{j=1}^k WIP_j}{DGR} \quad k = \text{Anzahl Toolgruppen} \quad (4)$$

Abbildung 2 zeigt eine grafische Darstellung der WIP-Aggregation.

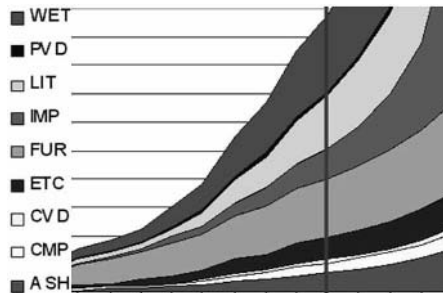


Fig. 2. WIP über der Auslastung.

2.3 Praxisrelevanz

Zweifelloos erfordert das vorgeschlagene Konzept der Multi-Toolgroup-OC einen höheren Aufwand in Datenerfassung und -verarbeitung. In der Regel sind die erforderlichen Quelldaten jedoch bereits vorhanden, da sie standardmäßig in der Betriebsdatenerfassung gespeichert werden. Es müssen meist einige Inkonsistenzen beseitigt werden, Berechnungsregeln formalisiert und automatisiert werden. Zudem ist eine Schnittstelle zum Reportingsystem zu gestalten. Interessanter Nebeneffekt ist, dass durch die Implementierung des Systems automatisch eine Konsistenz zu bereits vorhandenen oder aufzubauenden Toolgruppen- oder Workcenter-Betriebskennlinien entsteht. Diese Einzelkennlinien sind ein interessantes Instrument zur Analyse der Performance einzelner Fertigungsbereiche.

Das Konzept wurde in [6] untersucht und im Vergleich zu einem Simulationsmodell bewertet. Dabei konnte eine gute Realitätsnähe festgestellt werden.

3 Integratives Gesamtkonzept

Im Sinne eines Corporate Performance Managements sollte die Betriebskennlinien nicht isoliert betrachtet werden. Vielmehr ergeben sich eine große Zahl relevanter Schnittstellen zu anderen betrieblichen Systemen. Anwendungen im Fertigungsmanagement sind:

- Bottom up Kapazitätsplanung
- Top down Investitionsplanung
- Operative Liniensteuerung

Nicht zu vernachlässigen ist dabei der Aspekt Prozessperformanceanalyse innerhalb der Supply Chain. Letztlich ist die zu analysierende Durchlaufzeit der entscheidende Parameter für die Lieferfähigkeit des Produktionssystems. Daneben reflektiert die Variabilität des Systems, ausgedrückt durch das Alpha einer Toolgruppe bzw. durch die Form der BKL der Gesamlinie, die Liefertreue.

Typische Adressaten der Analyseergebnisse sind daher:

- Management
- Supply Chain Management / Logistik
- Line Control und Produktionssteuerung
- Industrial Engineering

Abbildung 3 verdeutlicht die Wechselwirkung des beschriebenen Modells innerhalb des betrieblichen Controllingsystems:

Neben der Analyse des Ist-Zustandes ist das vorgeschlagene realitätsnahe Betrieb-

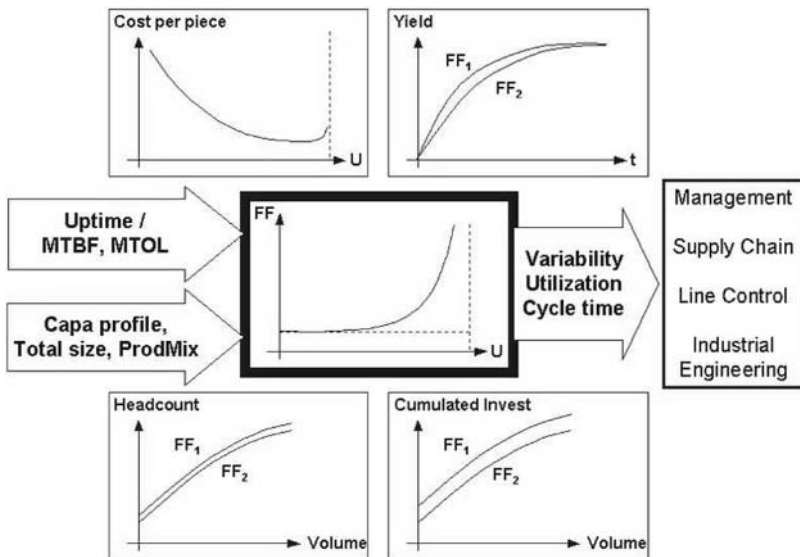


Fig. 3. Schnittstellen zur Betriebskennlinie.

skenlinienkonzept auch für Abschätzungen des Gesamtlinienvhaltens bei bestimmten Maßnahmen oder erwarteten Ereignissen geeignet. So wird in [1] der Einsatz von Workcenter-Betriebskennlinien zur Abschätzung des Systemverhaltens auf

ungleichmäßig verteilte Bestände im System vorgeschlagen. Eine weitere interessante Analysemöglichkeit bietet sich über die Option den Auslastungsgrad einzelner Toolgruppen planerisch zu ändern. Über die Verknüpfung der Toolgruppenauslastung mit der Maschinenzahl je Gruppe kann auf einfache Weise geprüft werden, welchen Einfluss Investitionen oder Deinvestitionen auf die Gesamtdurchlaufzeit der Fertigungslinie haben. Über eine Verknüpfung mit den Investitionsbeträgen je Anlage kann ein investioptimiertes Linienprofil zur Erreichung bestimmter Durchlaufzeitziele konstruiert werden. Diese Überlegungen sind natürlich mit einem gewissen Fehlerbalken behaftet. Dennoch können innerhalb kürzester Zeit verschiedene Thesen oder Ideen bewertet werden, bevor eine deutlich aufwändigere Simulation ausgewählte Vorschläge bestätigt. Zusammenfassend lässt sich sagen, dass das vorgeschlagene System der Multi-Toolgroup-OC bei minimalem zusätzlichen Aufwand eine Vielzahl von Einsatzmöglichkeiten bietet. Die Schwächen des bekannten einfachen BKL-Modells werden überwunden, sodass ein realitätsnahes Abbild der Fertigungsperformance entsteht, welches sogar für planerische Aussagen herangezogen werden kann.

References

1. Beeg T (2004): Wafer Fab Cycle Time Forecast under Changing Loading Situations, In: Proceedings of the Fifteenth Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference, Boston, MA, U.S.A., Piscataway, N.J.: IEEE ; San Jose, Calif., 2004.
2. Eichhorn D (2006): Controllingmodell zur Transformation von strategischen Unternehmenszielen in Fertigungsziele, Peter Lang, Frankfurt / Main.
3. Eichhorn D (2006): Process Measuring of Complex Manufacturing Processes, In: Nissen, Volker / Schneider, Herfried (Hrsg.): Prozessorientiertes Corporate Performance Measurement, Ilmenau 2006, S. 95-108.
4. Engelhardt C (2000) Betriebskennlinien: Produktivität steigern in der Fertigung. Carl Hanser Verlag, München u. a., 2000.
5. Hopp JW, Spearman ML (2000) "Factory Physics", Irwin: Chicago 2nd ed.
6. Obermaier G (2006): Vergleich verschiedener Ansätze zur Ermittlung von Betriebskennlinien am Beispiel der Halbleiterindustrie, Diplomarbeit TU München 2005.
7. Sakasegawa H (1977) "An Approximation Formula $L_q = \alpha \rho \beta / (1 - \rho)$ ", Annals Inst. Statist. Math., Vol. 29, Part A, 1977, pp. 67 - 75.

Über verschiedene Ansätze zur Ermittlung von Betriebskennlinien – Eine Anwendungsstudie aus der Halbleiterindustrie

Alexander Schömig¹, Dirk Eichhorn² and Georg Obermaier³

¹ Infineon Technologies AG, München, Germany

Alexander.Schoemig@infineon.com

² Qimonda AG, München, Germany

Dirk.Eichhorn@qimonda.com

³ Institut für Mikrostrukturtechnik, Forschungszentrum Karlsruhe GmbH,
Germany

Georg.Obermaier@imt.fzk.de

1 Einleitung

Der Bedarf an Modellen zur realitätsnahen Abbildung von Fertigungsprozessen ist nach wie vor aktuell. Gerade für das Verständnis fundamentaler Zusammenhänge in der Produktion sind geeignete Modelle von unschätzbarem Wert, denn logistische Prozesse sind in der Praxis oft schwer nachzuvollziehen, so dass für die Beschreibung des Verhaltens bestehender Systeme flexible und transparente Methoden gefordert werden, die mit einem vertretbaren Ermittlungsaufwand einher gehen. Kaum ein anderer Industriezweig als die Halbleiterbranche ist durch die verschärften Rahmenbedingungen eines überaus komplexen, kapital- und zeitintensiven Produktionsprozesses gekennzeichnet. Eingebettet in einem wechselhaften Umfeld nimmt die Prämisse des wirtschaftlichen Produzierens einen zentralen Stellenwert ein, so dass nicht nur marktseitige Forderungen mit den betriebsseitigen Bedingungen in Einklang gebracht werden müssen, sondern vor allem auch ein Konsens zwischen den betrieblichen Anforderungen gefunden werden muss. Die typischen Charakteristika der Mikrochipfertigung hinsichtlich Fertigungslogistik, Materialfluss in einem Fertigungsverbund und die Herausforderungen bei der Modellierung von Arbeitsschritten zur Lösung verschiedener planerischer und operationeller Probleme werden in [9] ausführlich dargestellt.

2 Grundlegende Ansätze und Ermittlungsverfahren

Die Betriebskennlinie stellt ein geeignetes Instrument dar, das den oben genannten Anforderungen gerecht wird. Dank ihres aussagekräftigen und anschaulichen Abbildungscharakters dient sie als ein Unterstützungswerkzeug im Bereich des taktischen

wie operativen Managements, um als Planungs-, Steuerungs- und Kontrollinstrument das Unternehmen zu führen. Die Betriebskennlinie (BKL), in der angelsächsischen Literatur oft auch als „Operating Curve“ [1], „Performance Curve“ [7] oder als „Characteristic Curve“ [4] bekannt, kann allgemein als ein Werkzeug zur Abbildung von Produktivitätskennzahlen verstanden werden. Dabei werden funktionale Zusammenhänge zwischen logistischen Kenngrößen wie Durchlaufzeit, Durchsatz und Bestand als deren klassischen Vertreter veranschaulicht. Ihre Objektivität ist folglich eher deskriptiver als präskriptiver Natur und trägt zum Gesamtverständnis eines gegebenen Systems bei. Das abzubildende System reicht von der Darstellung einer einzelnen Produktionsstation über eine Produktionsgruppe bis hin zu einer Produktionslinie und kann letztendlich auch im Produktionsverbund eingesetzt werden.

Zur Ermittlung einer Betriebskennlinie sind mehrere unterschiedliche Verfahren bekannt, die sich maßgeblich über ihre Beschreibungs- und Erklärungs-funktion im Hinblick auf die abzubildenden betrieblichen Prozessabläufe differenzieren. Es lassen sich die folgenden vier Ansatzmöglichkeiten kategorisieren, die zusammenfassend in Schaubild 1 wiedergegeben sind: Reale und ideale Betriebskennlinien, analytische und simulierte Betriebskennlinien.

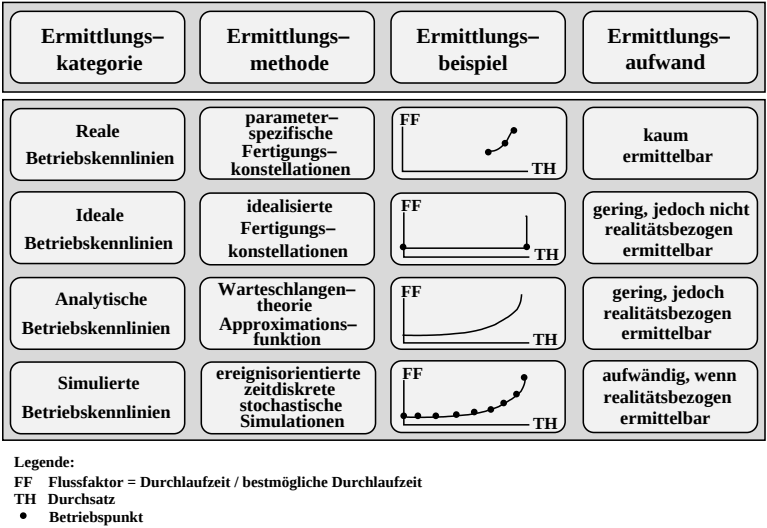


Fig. 1. Grundlegende Ansätze zur Ermittlung von Betriebskennlinien.

Auf Grund der enormen Vielfältigkeit von Parametereinstellungen und Produktionsszenarien ist die Gestaltung von realzeitgerechten Kennlinien eine äußerst schwer umsetzbare Aufgabe, die aus Zeit-, Verwaltungs- und Kostengründen nicht empfehlenswert ist. Idealisierte BKLs tragen zum Abbilden des gegenwärtigen Produktionsprozesses nicht befriedigend bei, da unter der Annahme idealer Rahmenbedingungen der reale Bezug unterbleibt. Sie stellen jedoch als „best-case-Szenario“ einen wichtigen Ausgangs- und Orientierungspunkt dar. Folglich sind analytische Methoden und ereignisorientierte zeitdiskrete stochastische Simulationsrechnungen

in der Praxis die beherrschenden technischen Mittel, weil diese mit der Berücksichtigung der spezifischen Produktionsmerkmale umfassender einhergehen können. Bezugnehmend auf die Unterschiede zwischen den genannten Ansätzen kann festgehalten werden, dass unter Voraussetzung eines bestehenden analytischen Modells sich die Kennlinien aufwandsärmer nachbilden lassen und funktionale Zusammenhänge zwischen zwei Systemgrößen direkt aus diesem sich ableiten lassen. Im Hinblick auf die Vorteilhaftigkeit der schnellen Ermittlung von Leistungskenngrößen anhand einfacher quantitativer Zusammenhänge sind diese einem simulativen Experiment vorzuziehen. Im weiteren Verlauf der Studie wurden daher primär analytische Methoden kritisch diskutiert.

Die analytischen Ermittlungsmethoden lassen sich in die beiden deduktiven Teilbereiche der Warteschlangentheorie und der ableitbaren Approximationsfunktionen untergliedern:

• Warteschlangentheorie

Hierunter fallen Ansätze, die sich auf aus der Warteschlangentheorie, in der Regel der sog. Kingman-Gleichung [6], abgeleitete Formeln stützen, sowie deren Erweiterungen auf Mehrbedienersysteme und Anpassungen für den Einsatz in der Praxis. (Siehe hierzu [3] und [5], bzw. den Referenzen hierin.)

$$FF = \frac{c_a^2 + c_e^2}{2} \cdot \frac{U\sqrt{2(m+1)}-1}{m(1-U)} + 1 \quad (1)$$

FF = Flussfaktor; c_a, c_e = Variationskoeffizienten;

U = Auslastung; m = Maschinenanzahl

Diese Formel wird genutzt zur Berechnung von Betriebskennlinien für eine Einzelanlage, einer Anlagengruppe von zumindest nahezu identischen Anlagen, bis hin zur gesamten Fertigungslinie. In letzterem Fall sind umfangreiche vereinfachende Annahmen zur Aggregation von Parametern notwendig, deren Wichtigste die Bestimmung der Gesamtlinenkapazität bzw. Auslastung darstellt, wie in [2] beschrieben.

• Empirisch abgeleitete Approximationsfunktionen

Der vornehmlichste Vertreter ist hier eine auf dem sog. "Trichtermodell" und dem Durchlaufdiagramm basierende Beschreibung und Ableitung elementarer Zusammenhänge verfahrenstechnischer Fließprozesse. Dies ist eine im Kern deduktiv-experimentelle Vorgehensweise, wobei hierfür allgemein eine Konzentration auf der Darstellung des idealen Leistungs-Bestands-Verhältnisses erfolgt. Dabei bestimmt die Kapazität die obere Leistungsgrenze des Systems. Ohne weitere mathematische Herleitung mündet diese Vorgehensweise in der sog. C_{Norm} -Funktion [8].

$$FG = 1 + \frac{\alpha_1 \cdot \sqrt[C]{1 - \left(1 - \frac{L_m}{L_{\max}}\right)^C}}{\frac{L_m}{L_{\max}}} \quad (2)$$

FG = Flussgrad; α_1 = Streckfaktor; $C = C_{\text{Norm}}$ -Wert;

L_m = mittlere Leistung; $L_{\max}$ = maximale Leistung

Werden diese beiden Ansatzmöglichkeiten miteinander verglichen, so unterstreicht Abbildung 2 die Erkenntnis, dass die beiden Funktionsgleichungen in keiner Äquivalenzbeziehung zueinander stehen. Begründet wird dies über die unterschiedlichen Divisoren der jeweiligen Auslastungsterme.

Warteschlangentheorie	Ableitbare Approximationsfunktion
$FF = \underbrace{\alpha}_{\text{Variabilität}} \cdot \underbrace{\frac{\frac{TH}{CAP}}{1 - \frac{TH}{CAP}}}_{\text{Auslastungsterm}} + 1$ $\lim_{TH \rightarrow 0} FF = 1$ $\lim_{TH \rightarrow CAP} FF = \infty$	$FF = \underbrace{\alpha_1}_{\text{Streckfaktor}} \cdot \underbrace{\frac{\sqrt[C]{1 - (1 - \frac{TH}{CAP})^C}}{\frac{TH}{CAP}}}_{\text{Auslastungsterm}} + 1$ $\lim_{TH \rightarrow 0} FF = 1$ $\lim_{TH \rightarrow CAP} FF = \alpha_1 + 1$

Fig. 2. Analytische Ermittlungsansätze im Vergleich.

3 Experimentelle Studien

Zum Vergleich der verschiedenen Ansätze wurden verschiedene Vergleichsstudien anhand von Beispielmодellen durchgeführt. Abbildung 3 zeigt unterschiedlich generierte Betriebskennlinien mittels des Simulationsmodells Nr. 1 nach MIMAC. Hierbei wird angenommen, dass das Simulationsmodell die Daten generiert, wie sie in der modellierten Realität auftreten würden und passt dann gemäß der gewählten Verfahren entsprechend die Modellparameter an. Dies geschieht in der Regel anhand eines bestimmten Betriebspunktes.

Es wird verdeutlicht, wie differenzierend die Performance einer Produktionslinie verstanden werden kann. Dabei wird das genaue Interpretationsverständnis im nächsten Kapitel dargelegt.

4 Zusammenfassung und Schlussfolgerungen

Sowohl die berechneten Kennlinien auf Basis der Warteschlangentheorie, als auch auf Basis der ableitbaren Approximationsfunktionen stellen in sich einfache und leicht nachvollziehbare Modelle dar. Für die Betriebskennlinienansätze auf Basis der Warteschlangentheorie ist der Ansatz der Identifizierung der Engpassgrenze auf Basis der mittleren Linienkapazität nicht zu empfehlen, da unter Umständen die durchsatzlimitierende Kapazitätsgrenze der Produktionslinie überschritten wird. Als ‚performance constraint‘ einer Produktionslinie eignet sich die Identifizierung über das Anlagensystem, das die minimale Effizienz einbringt. Weiterhin ist bei allen Ansätzen auf Basis der Warteschlangentheorie zu beobachten, dass keine Kapazitätsvollauslastung herbeigeführt werden kann, aber eine Integrationsflexibilität bezüglich weiterer Produktionsleistungskennzahlen zu verzeichnen ist. Der gegebene Vorteil der individuellen Berücksichtigung der Variabilitätseinflüsse der einzelnen Maschinenanlagen wird jedoch nur mit der Aggregationstechnik der ”Multi-Tool-group Betriebskennlinie” nach Eichhorn [2] Rechnung getragen und lässt die Ermittlungsgenauigkeit gänzlich steigen. Allerdings ist dies mit einem deutlich erhöhten Ermittlungsaufwand verbunden, da alle am Fertigungsprozess beteiligten Produktionssysteme berücksichtigt werden müssen.

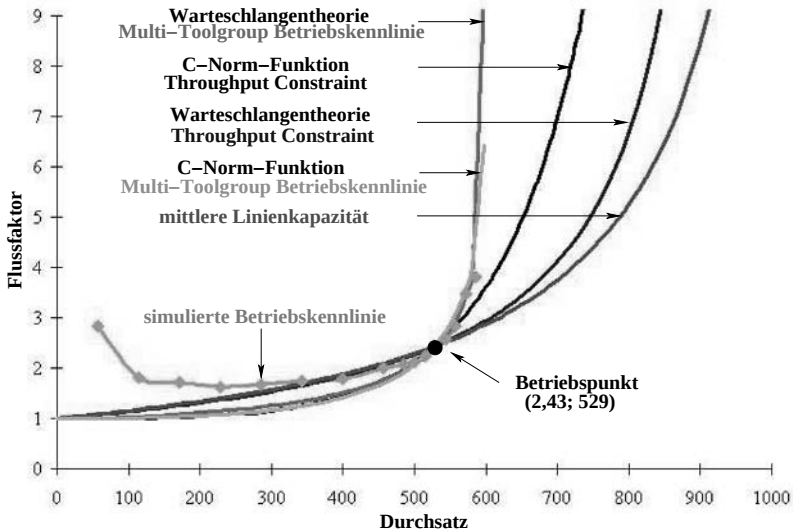


Fig. 3. Flussfaktor-Durchsatz-Betriebskennlinie am Beispiel des Simulationsmodells nach MIMAC [4].

Für die abgeleitete Approximationsgleichung auf Basis der C_{Norm} -Funktion ist die Möglichkeit der Erreichung der Kapazitätsgrenze als durchaus positiv zu bewerten. Hingegen ist die Integration weiterer Produktionsleistungskennzahlen, zumindest auf Basis der vorgeschlagenen Grundgleichung problematisch, was die Flexibilitätseigenschaft stark einschränkt. Zwar sind gut geführte Funktionsbibliotheken bei der Suche nach neuen Grafen zur Beschreibung des abzubildenden Prozesscharakters hilfreich, die durchzuführenden Transformationsschritte jedoch aufwändig, da sich hier im Gegensatz zur Warteschlangentheorie die wissenschaftliche Forschung noch in einem Anfangsstadium befindet und auf fertige problemspezifische Modelle nicht zurückgegriffen werden kann.

Wird aber auf die Berücksichtigung weiterer Einflussparameter kein großer Wert gelegt, so ist der Ermittlungsaufwand natürlich gering, da die vorgestellten Gleichungen umgänglich und überschaubar sind. Da gezeigt wurde, dass auch hier die Aggregationstechnik zur Ermittlung der BKL einer Produktionslinie Verwendung finden kann, ist der Aufwand im Vergleich zum Ausgangsmodell nach [10] etwas höher einzustufen, gewinnt zugleich aber an deutlicher Ermittlungsgenauigkeit.

Gerade aufgrund der mangelhaften Integrationsflexibilität erscheint es am geeigneten, die Verfahrensmethodik der "Multi-Toolgroup Betriebskennlinie", generiert auf Basis der Warteschlangentheorie, als das momentan adäquateste Instrument zur Beschreibung verfahrenstechnischer Fließprozesse zu bezeichnen. Auch ist die Grundgleichung im Hinblick auf den praktischen Einsatz noch überschaubarer als die der C_{Norm} -Funktion. Interessant wäre es aber weiterhin, das resultierende Kennlinienverhalten der auf Basis der vorgeschlagenen Ausbeuteintegrationsfunktion der ableitbaren Approximationsgleichung mit den Ergebnissen der Warteschlangentheorie weiterführend zu untersuchen.

References

1. Aurand SS, Miller PJ (1997) The operating curve: a method to measure and benchmark manufacturing line productivity. IEEE / SEMI Advanced Semiconductor Manufacturing Conference and Workshop (ASMC) 391-397
2. Eichhorn D (2005) Controllingmodell zur Transformation von strategischen Unternehmenszielen in Fertigungsziele. Peter Lang, Frankfurt / Main
3. Engelhardt C (2000) Betriebskennlinien: Produktivität steigern in der Fertigung. Carl Hanser, München
4. Fowler JW, Robinson JK (1995) Measurement and improvement of manufacturing capacity (MIMAC) project report. SEMATECH Technology Transfer 95062861A-TR
5. Hopp JW, Spearman ML (2000) Factory physics: foundations of manufacturing management, 2. Aufl. Irwin McGraw-Hill, Boston
6. Kingman JFC (1961) The single server queue in heavy traffic. Proceedings of the Cambridge Philosophical Society 57:902-904
7. Martin DP (1999) Total operational efficiency (TOE): the determination of 2 capacity and cycle time components and their relationship to productivity improvements in a semiconductor manufacturing line. IEEE / SEMI Advanced Semiconductor Manufacturing Conference and Workshop (ASMC) 380-385
8. Nyhuis P, Wiendahl H-P (2003) Logistische Kennlinien, 2. Aufl. Springer, Berlin Heidelberg New York
9. Schömig A (2000) OR Probleme in der Mikrochipfertigung. Proceedings Symposium on Operations Research 1-6
10. Wiendahl, H-P, Schneider M (2001) Entwicklung von Logistischen Kennlinien für Fertigungsbereiche. PPS Management 6, 2: 9-13

The Use of Chance Constrained Programming for Disassemble-to-Order Problems with Stochastic Yields

Ian M. Langella and Rainer Kleber

Faculty of Economics and Management, Otto-von-Guericke University Magdeburg, Germany

ian.langella@ww.uni-magdeburg.de

rainer.kleber@ww.uni-magdeburg.de

Summary. Stochastic yields from disassembly complicate the planning of so-called disassemble to order problems, where a specified amount of components must be harvested from various models of returned products. Chance constraint programming, a branch of stochastic programming, has proven useful in several applications of operations management. This contribution will first formulate a novel chance constrained programming model for the single-period disassemble-to-order problem. We will then illustrate its application using an example, and highlight the tradeoff between service and costs which emerges. We also suggest a variety of extensions to the basic model, many of which will likely prove to be trivial and relevant to industry.

1 Introduction

Within the realm of closed loop supply chain management, products which are no longer needed by customers are collected and transported back to the manufacturer [1]. The firm receiving these returned products (called *cores*) can either recover material through recycling, remanufacture the products, or otherwise properly dispose of the cores. Remanufacturing cores entails disassembling the core completely, cleaning and inspecting the parts, adding newly produced parts, and reassembling the product to meet “good as new” quality standards. This option is particularly advantageous when there is a market for remanufactured products, as seen in e.g. automotive parts and photocopiers.

Planning remanufacturing starts by converting the demand for remanufactured models to a demand for components based on an MRP explosion. Given demand for components, it must then be decided how many cores to disassemble in order to fulfill this demand. This is complicated by the uncertainty surrounding the quality of the returned products, which manifests itself in random component-specific yields from disassembly [2]. This yield uncertainty is such that using the mean yield rate (and

thereby treating the problem as deterministic) often fails to provide an adequate number of components.

Two well-known alternative stochastic programming approaches are recourse models and chance constraint programming (CCP) models. In recourse models, decisions are first separated into first stage (made before the revelation of the random variable, here the yield) and second stage (made after the random variable is realized). Thus the second stage recourse decisions (which are weighted by penalty costs in the objective function) compensate for the effects of the randomness and ensure the feasibility of constraints [3]. The disadvantage of recourse models is that (1) quantifying these recourse costs is often quite difficult and (2) these models are difficult to solve for realistic sized problems. In certain situations however, it suffices that the constraints hold with a certain probability (practical examples might include e.g. emergency services or power system generation planning). Under CCP and in contrast to recourse models, constraints cannot be fulfilled in any case (there is no recourse) and therefore we seek to define an acceptable amount of risk that the constraints are violated [3].

This work will provide an initial look into how CCP models might be employed in disassembly planning under stochastic yields. In the next section, we will put forth the model and demonstrate its application through a numerical example paying particular attention to the tradeoff between service and cost. In the final section, we will provide some concluding remarks and suggest extensions which are of special interest from a practical point of view.

2 Model

In this section, we will use CCP to model a disassemble-to-order problem with stochastic yields. We assume (1) that cores are completely disassembled, (2) that there is no restriction on either the amount of cores which can be obtained from the market or (3) the disassembly capacity, (4) that disposal costs are negligible, (5) that demand for leaves is deterministic, (6) that the firm has no ability to procure leaves as an alternative to disassembly, and (7) that management has defined a service level (a prescribed probability that the demand is fulfilled) for each leaf. Additionally, for this initial work we will restrict our attention to a single planning period. While these assumptions are somewhat restrictive, many enjoy substantial empirical support. In the authors' experience, many products are completely disassembled prior to remanufacturing, among them being automotive engines. Superfluous leaves are often recycled for material rather than being landfilled, and as such generate a marginal amount of revenue, which motivates the assumption of negligible disposal costs. Although the assumption of deterministic demand is easy to relax, it is often the case in practice that demand is specified by another division, and can therefore be treated as given. This leaves yield uncertainty as the only source of stochastic influences. The ability to procure parts to be used in remanufacturing either from serial in-house production (which is particularly advantageous when being produced in mass) or through an external supplier might not be possible later in the life cycle as mass production has stopped and suppliers might be uninterested in providing relatively small quantities to the firm (or might only provide such parts at a high cost and/or long lead times). We will later discuss the case when lead times are

long, but for the moment assume that the firm has no access (internal or external) to leaves. Relaxing the assumptions that there are no limits on the amount of cores which can be procured (constrained returns) or disassembled (constrained disassembly capacity) will also prove trivial. Lastly, service levels for leaves would be a managerial decision, albeit a particularly difficult one to make and one with ramifications on costs as we will soon see.

We can start defining notation by introducing $i \in I$ our core index and set and $k \in K$ for leaves. The cost to acquire and transport cores to the disassembly facility, completely disassemble the core, and clean and inspect the leaves (referred to as *core cost*) is denoted c_i^z . The disassembly of a certain core i results in harvesting π_{ik} units of leaf k (the yield of disassembly), which is stochastic. As in prior research, we model yield uncertainty using the technique of stochastically proportional yield which fits well with standard CCP formulation, while quickly noting that this is by no means a prerequisite for employing CCP. The specified demand per leaf is denoted D_k . A service level for each leaf is denoted α_k . Our decision variable x_i^z represents the amount of core i which should be disassembled.

The objective function and constraints can be put forth as:

$$\min \quad C = \sum_{i \in I} c_i^z \cdot x_i^z \quad (1)$$

$$\text{s.t.} \quad \Pr \left\{ \sum_{i \in I} x_i^z \cdot \pi_{ik} \geq D_k \right\} \geq \alpha_k \quad \forall k \in K \quad (2)$$

$$x_i^z \geq 0 \text{ and integer} \quad \forall i \in I \quad (3)$$

As can be seen, the objective function (1) minimizes the sum of the core cost and the chance constraints (2) prescribe that the demand is fulfilled with a minimum probability of α_k . If the yields were deterministic, the term within brackets would constitute a hard constraint while the objective would remain unchanged. Lastly, (3) requires that the core decisions be non-negative integers. The model seeks to decide on the right mix of cores in order to fulfill the constraints with their specified probabilities.

Without going into too much detail on how one solves CCP models (the interested reader can be referred to [1] for an approachable starting reference and [4] for a more advanced treatment) the stochastic constraints are converted into deterministic equivalents which in the most simple case would result in an LP formulation. In order to simplify things, it is helpful to assume that the random variables are normally distributed and independent of each other, leading to a convex optimization problem. In such a case, for each random yield variable only the mean (μ_{ik}^π) and standard deviation (σ_{ik}^π) are required. If the random variables were not independent their covariance would have to be incorporated, which is more complicated but still possible.

Example. In order to illustrate the model and its application, this section will provide a numerical example. In this example, we will take a look at the effect of the α -values on the decisions and resulting cost by examining solutions for various levels. The product structure for the numerical example is given in Figure 1, which consists of a two-core, three-leaf structure with one leaf (B) common to both cores. Remaining parameters are given in Table 1. This instance is typified by symmetric core costs and demands, and a yield distribution with a large dispersion (while having

a minimal probability mass falling outside the range $[0,1]$). Under such symmetric data, it will be the case that the disassembly decisions for both cores will be identical.

Table 1. Parameters for example

Parameter	Value
Core cost	$c_1^z = c_2^z = 10$
Leaf demands	$D_1 = D_3 = 100, D_2 = 200$
Yield distribution moments	$\mu_{ik}^\pi = 0.5, \sigma_{ik}^\pi = 0.15 \forall i, k$



Fig. 1. Product structure for numerical example

The results (decisions and costs) for various α -values are given in Table 2. As can be seen, if we were willing to fulfill the demand only with a probability of 0.5, we would merely disassemble enough cores (200 of each) to meet demand based on the average yield value. This is analogous to a newsvendor problem from standard inventory control with a service level of 50%. As our α -values increase, we require additional safety (i.e. additional amounts of cores to disassemble) in order to meet the demand with the specified probability. For instance at the level of $\alpha = 0.95$, the solution requires the disassembly of roughly double the amount of cores as the solution for $\alpha = 0.5$. Figure 2 depicts the increase in decisions with an increasing α . As can be seen, for $\alpha \geq 0.95$, the costs increase particularly rapidly. Of course, this would be less drastic for yield distributions with smaller variance than the one used in this example.

Table 2. Decisions and costs for various levels of α

α	x_1^z, x_2^z	C	$\hat{\alpha}_A, \hat{\alpha}_C$	$\hat{\alpha}_B$
0.5	200	4000	0.5	0.5
0.6	217	4340	0.6	0.64
0.7	238	4760	0.7	0.77
0.8	268	5360	0.8	0.88
0.9	325	6500	0.9	0.96
0.95	395	7900	0.95	0.99
0.99	663	13260	0.99	> 0.999

This brings us to an interesting point, which can be illustrated by the example and might be presumed to hold more generally. In our example, symmetric demands

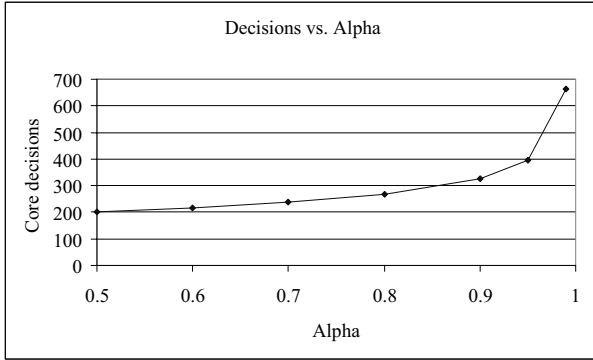


Fig. 2. Graph depicting core decisions as alpha value grows

and yields was combined with the structure where each core had one unique leaf (A in 1 and C in 2) and one common leaf (B). We can also see in Table 2 actual service levels (denoted by $\hat{\alpha}_k$) provided by the solution. As we can see, in each case (other than when $\alpha = 0.5$) the solution is such that only the chance constraints for the unique leaves are binding, providing a higher probability of demand fulfillment to the common leaf. This is naturally due to risk pooling in this instance, but it also shows us that demand for unique leaves lead to lower bounds on the disassembly decision for the respective core. These bounds are easy to be calculated and can be used in a solution procedure. In our example for instance, the solution was completely determined by these bounds. Risk pooling effects but also integrality constraints lead to actual service levels being higher than prescribed minimum levels.

3 Conclusion and Outlook

This contribution has provided a novel CCP model to be used in the disassemble-to-order problem with stochastic yields and we have seen in an example how providing higher service brings about additional disassembly costs. The example also illustrated an effect of risk pooling in the yield of leaves which are common to more than one core. This model can be extended in a plethora of directions, many of which will be of high interest to practitioners. Demand randomness can be considered in much the same manner as we have seen with yield uncertainty. A return constraint can easily be added, limiting the amount of cores which can be obtained from the market prior to being disassembled. Indeed, CCP can easily handle also a stochastic return situation where the amount of returns is random with a known mean and variance. The model can be modified to account for disassembly capacities, even in cases where the amount of capacity and/or absorption rates are stochastic. In incorporating either a disassembly capacity or return constraint it will likely be necessary to add an additional sourcing option for leaves to guarantee a feasible solution for a given service level. To do this, the procurement of leaves can be included by adding applicable terms to the objective function and chance constraint. This might prove particularly useful in cases where the procurement decision (due to long lead times, as is often the case in several practical settings)

must be made in advance of the disassembly decision. These preceding extensions will in all likeliness turn out to be trivial. More difficult extensions, which will be the subject of future work, would be to extend the framework beyond the single period setting presented here and moreover to examine how one should set α -values for leaves. Finally, our model dealt with individual chance constraints for each leaf. For the later reassembly, however, it would be more important to have all relevant leaves available for the remanufactured product. This would lead to the necessity of incorporating joint chance constraints, and will be therefore relegated to future work.

Acknowledgements. We express our thanks to Stefan Minner, who suggested this research and contributed much with his typically helpful and enlightening comments.

References

1. de Brito MP, Dekker R (2004) A framework for reverse logistics. In: Dekker R, Fleischmann M, Inderfurth K, van Wassenhove LN (eds) *Quantitative Approaches to Reverse Logistics*. Springer, Berlin
2. Inderfurth K, Langella IM (2006) Heuristics for solving disassemble-to-order problems with stochastic yields. *OR Spectrum* 28:73-99
3. Sen S, Hingle JL (1999) An introductory tutorial on stochastic linear programming. *Interfaces* 29:33-61
4. Charnes A, Cooper WW (1959) Chance constrained programming. *Management Science* 6:73-79
5. Taha HA (2003) *Operations research: an introduction*. Seventh Edition. Prentice Hall, Upper Saddle River.
6. Prekopa A (1995) *Stochastic programming*. Kluwer, Dordrecht.

Optimal Usage of Flexibility Instruments in Automotive Plants

Gazi Askar¹ and Jürgen Zimmermann²

¹ DaimlerChrysler AG Group Research, Ulm, Germany
`gazi.askar@daimlerchrysler.com`

² Department for Operations Research, Clausthal University of Technology,
Germany
`juergen.zimmermann@tu-clausthal.de`

Summary. Scope of this work is the development of an optimization approach to find the optimal configuration of automotive plant capacities for a time horizon of one to seven years regarding market demand, planning premises, and dependencies between the shops (body shop, paint shop, final assembly) and between planning periods (e.g. learning curves). The methodology has been implemented in a planning tool and includes production and workforce planning. The optimization algorithm is based on a dynamic programming approach.

1 Introduction

In the past years the automotive market is characterized by increasing diversification and shortened lifecycles, which leads to an enhanced fluctuation of demand. This trend is enforced by seasonal effects and a rising uncertainty in demand forecast. In order to achieve a high utilization OEMs implement a high degree of flexibility in their plants. Many approaches for planning and optimizing the degree and structure of flexibility in production facilities have been considered (see for example [6], [5]). The next nontrivial but important planning task is the optimal use of flexibility instruments for adapting capacities to demand. In particular, industries with a high proportion of personnel costs like the automotive sector use organizational flexibility instruments just as technical flexibility so that these two aspects can not be treated separately or successively. Regarding these requirements we emphasize an integration of workforce and production planning with a focus on changing demands using given flexibility instruments (see Fig. 1).

The simultaneous medium-term planning of production and workforce capacities is not self-evident. The subject first arose in the late 80s when the general economic situation led to a broad public and scientific discussion of the topic. On the scientific side pioneering work was conducted by [4] and [3]. In practice the combination of production and workforce planning is rare, although there is a urgent need to manage available flexibility efficiently ([2], [7]). Due to special requirements regarding the

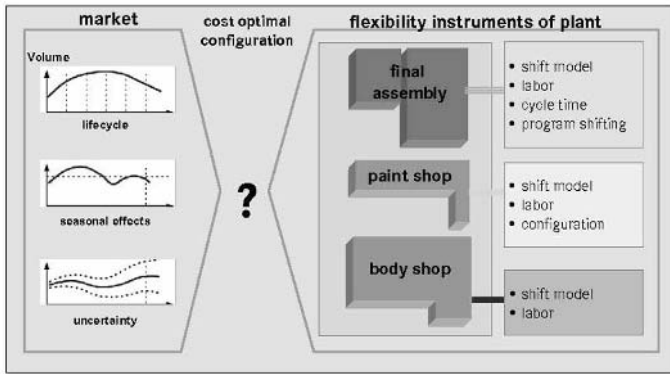


Fig. 1. Demand of flexibility and flexibility instruments

flexible workforce planning and an enormously rising complexity of the planning problem it is still not common that labor aspects are considered in the context of production planning.

2 Modeling an Automotive Plant

The manufacturing of a car is divided in three shops. The first step is welding the car body in the highly automated body shop. Workforce is mainly required for programming and maintenance of the machines. In the succeeding paint shop the surface of the car body is prepared and lacquered. In general, the painting process is fully automated. Only handling and rework is done manually. In the last step the coated car body is composed with all modules and parts in the final assembly. The low degree of automation leads to a high degree of flexibility in this shop, but also to a high integration of workforce and production planning. Hence, these two planning tasks have to be considered simultaneously together with planning of working time and an evaluation of costs in order to find an optimal adaptation of automotive plants to given demands. Additionally, there has to be a coordination among the different shops, which are uncoupled. The main functions of the buffers between succeeding shops are catching breakdowns and sorting the cars for the production process. Only a small part of each buffer remains for working with different production rates and shift lengths in successive shops.

As described above final assembly possesses most of flexibility instruments. The production is organized by an assembly line with constant speed. All assembling

operations are assigned to sections of same lengths, where each section is filled with at least one worker. The cycle time for each section is determined by the speed of the conveyor. Increasing the line speed implies the following three effects. Firstly, the amount of possible operations per section decreases, so that the amount of sections and with it the labor demand increases. Secondly changing the cycle time initiates a learning process in which the required time for each operation is increased. Thus, there is an additional labor demand during this time. Finally, changing speed causes costs for installing the required tools at the sections.

From the organizational point of view there are three main instruments for varying the available working time. The adaptation with the highest impact on the production capacity is changing the amount of shift groups (e.g. working in night shift). The next step would be changing the amount of planned shifts (e.g. Saturday shifts). Smallest adaptations can be done by changing the operating time per shift. Each of these organizational instruments have to be used under consideration of the total working hours. According to the labor contracts these can exceed the agreed working time, but only to a certain degree and for a certain period. In order to keep the given corridor for the total working hours, workforce capacity has to be increased or working time has to be reduced.

One additionally available instrument is shifting production volume to previous and succeeding periods in order to prevent expensive adaptations in periods with high production peaks. Usually the described planning task is accomplished for the next one to seven years, so that most of the demand is not fixed but forecasted and has to be negotiated with marketing and sales.

While the organizational instruments of paint and body shop are the same as in final assembly, the technical adaptation possibilities are different. Painting cars is accomplished in successive production steps with small buffers between the stations. Technical adaptations are done by enabling and disabling several stations in different production steps. Each possible combination of activated stations is defined as one configuration and represents one output rate. It can differ in each shift group. Changing the configuration influences the labor demand and causes change costs.

Typically, the body shop has no technical flexibility because the cycle time is fixed due to the available robots. Acquiring additional or faster machines would be possible, but is out of scope.³

The production program of paint and body shop has to be determined in combination with the production program of final assembly. It has to be ensured, that the buffer between two shops is always filled to the minimum inventory level. Unless the buffer available for asynchronous capacity strategies in successive shops is very small, as mentioned before, each shifting of production program would have a relatively direct effect on the neighboring shop.

3 Approach for Optimization

Based on the description above single line models for final assembly, paint shop, and body shop have been implemented. In order to find an optimal adaptation for an automotive plant the planning problem has been divided in two levels. On the lower level each production line in each shop is optimized individually, using a dynamic

³ This would be an issue for the strategic planning

programming approach (see [1]) and assuming that there are no dependencies between the lines in one shop. Due to the high complexity on plant level a successive optimization of each shop, starting with the final assembly, has been used to find a global solution.

The planning problem for one production line can be divided in periods, where each period has a certain amount of possible states. Each state represents a certain configuration of production and workforce capacity and the previous development of state variables depending on the particular path to the current state. Some planning aspects like the working time account or the learning curve over more than one period after a change of the line speed are contrary to the premise of dynamic programming that each state depends only on the state of and the decision made in the preceding period. Status variables [1] were used to solve this conflict. As a result the amount of states per period and the complexity of the problem increases. For the final assembly the state space SFA_t in each period is determined by

$$SFA_t = (ct_t, lc_t, wtm_t, sa_t, sc_t, wta_t, b_t)$$

ct_t :	cycle time
lc_t :	change of labor capacity
wtm_t :	working time model (shift groups and operating time)
sa_t :	planning of Saturday shifts
sc_t :	status variable for amount of periods since last cycle time change
wta_t :	status variable for total working hours
b_t :	status variable for accumulated program shifting since now

The amount of possible values for this parameters defined by boundaries and the discretization determines the complexity of the problem. The state spaces for a production lines in paint and body shop are similar.

In each period t a decision for the next period $t+1$ has to be made, which consists of a choice for the technical capacity ($\widehat{ct}_t$), the organizational capacity ($\widehat{wtm}_t, \widehat{sa}_t$), the labor capacity ($\widehat{lc}_t$), and the production program. The state in period $t+1$ depends on the preceding state and the decision that has been made. The cycle time and the working time model are defined directly by the decision. The labor capacity is calculated by the labor capacity of the preceding period t and the decision for hiring or dismissing worker. The difference between period demand and production volume represents the buffer level change in period $t+1$. The sum of the buffer level in t and the buffer level change in $t+1$ determines the state of the buffer level b_{t+1} . The total working hours are also calculated from the change of working hours in $t+1$ and the total working hours in t . The value of total working hours in $t+1$ defines the current state of wta_{t+1} . At least, the value of sc_t is set to zero if $ct_t \neq ct_{t+1}$ and is increased by one otherwise. Each state that does not satisfy the production or labor demand or violates the restrictions for cycle time changes or total working hours is invalid. For a more detailed description of this approach and the possibilities for reducing the complexity of the problem see [8].

The total labor costs (wages and additional payments, e.g. for night shifts) and the change costs (e.g. changing cycle time) are used for the evaluation of each state. The objective is to minimize the reduced sum of costs over all periods. Let T be the number of examined periods, pc_t the personal costs, ctc_t the costs for changing technical capacities, $wtmc_t$ the costs for changing organizational capacities, lcc_t the costs for changing labor capacity, as well as δ_t , ϵ_t , and ϕ_t binary variables indicating

some change, and i the interest rate, then the total costs over all periods are given by:

$$F^*(s_1, \dots, s_T) := \min \left(\sum_{t=1}^T (p_{ct} + \delta_t \cdot ct_{ct} + \epsilon_t \cdot wtm_{ct} + \phi_t \cdot lcc_t) \left(1 + \frac{i}{100}\right)^{-t} \right).$$

The planning task can be illustrated as a min-cost-path problem (see Fig. 2). Thereby the states correspond to nodes and the decisions to edges. Infeasible states and invalid decisions are eliminated dynamically.

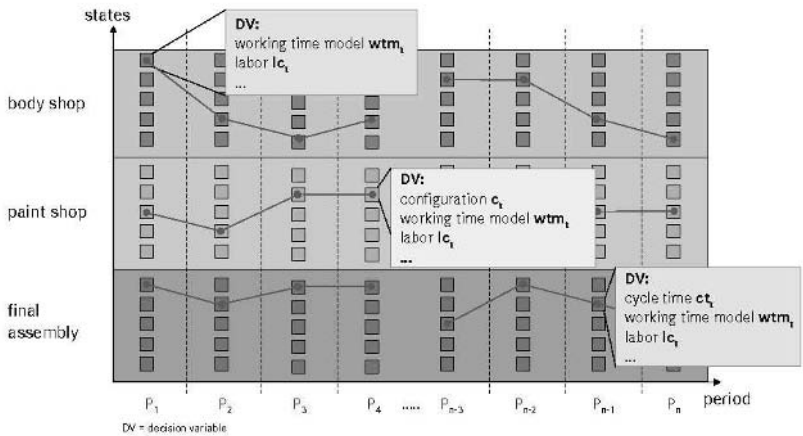


Fig. 2. Dynamic programming approach for solving single line planning problems

After all single line problems are solved to optimality in terms of the state space and the chosen discretization a coordination between the shops has to be found. The coordination between final assembly, paint shop and body shop is solved through a heuristic approach. An optimal configuration for the paint shop is searched under restriction of the optimal assembly configuration and the available buffer. That means, the valid state space for the preceding paint shop is reduced, because possible states in the paint shop with a high difference in capacity relating to final assembly become infeasible due to the small buffer size. After finding an optimal solution for the paint shop, the same proceeding is applied to the body shop.

4 Conclusion

The entire methodology has been implemented in form of a planning tool called "Lifecycle Adaptation Planner (LAP)" that has been transferred to and is in use in several DaimlerChrysler plants in Germany and in other countries. It was shown that the optimization of real world problems could be performed in an appropriate time between a few seconds and a few hours, depending on the specific problem premises. Often additional restrictions (e.g. given by management) can be added that help to shorten the required time to at most one hour.

References

1. Bertsekas DP (2000) Dynamic Programming and Optimal Control. Athena Scientific, Belmont
2. Dechow M, Eichstädt T, Hüttmann A, Kloss M, Mueller-Oerlinghausen J (2005) Reducing Labor Costs by Managing Flexibility. McKinsey Company Automotive Assembly Extranet
3. Faißt J, Schneeweiß C, Wild B (1992) Flexibilisierung der Personalkapazität durch zeitliche und räumliche Umverteilung. In: Kapazitätsorientiertes Arbeitszeitmanagement. Physica, Heidelberg
4. Günther HO (1989) Produktionsplanung bei flexibler Personalkapazität. Poeschel, Stuttgart
5. Pibernik R (2001) Flexibilitätsplanung in Wertschöpfungsnetzwerken. Deutscher Universitätsverlag, Frankfurt (Main)
6. Schneeweiß C (2002) Einführung in die Produktionswirtschaft. Springer, Berlin
7. Scherf B (2004) Ganzheitliches Ressourcenmanagement erfordert Zusammenspiel von PPS und Personaleinsatzplanung. In: (4) PPS Management. Gito Verlag, Berlin
8. Zimmermann J, Hemig C, Askar G (2006) Ein Dynamischer Optimierungsansatz für die Montageplanung. In: Wirtschaftswissenschaftliche Schriftenreihe TU-Clausthal, WIWI-Report No. 6

Comparison of Stochastic- and Guaranteed-Service Approaches to Safety Stock Optimization in Supply Chains

Steffen Klosterhalfen and Stefan Minner

University of Mannheim, Schloss, 68131 Mannheim, Germany
`klosterhalfenbwl.uni-mannheim.de`
`minner@bwl.uni-mannheim.de`

Summary. Contributions to multi-echelon safety stock optimization can be classified as being either of the stochastic- or guaranteed-service framework. Though both approaches have been used in inventory theory for a long time, there is still a lack of a detailed comparison. In this paper, such a comparison is presented. Differences in the materials flow are outlined and insights into the performance of both approaches are gained from a simulation study.

1 Introduction

In supply chains, safety stock is a prominent means to deal with demand uncertainty. Developing computationally tractable approaches for determining the appropriate amount of safety stock in fairly general systems is a complex task. Models that deal with the optimization of stock levels in a multi-echelon context fall into two major categories (see, e.g., [3]): stochastic-service approach (SSA) and guaranteed-service approach (GSA) models. The approaches differ in terms of the underlying materials flow concept and the resulting service time characteristics. The service time is the time it takes until an order placed by a stockpoint is shipped to the stockpoint by its predecessor. In the SSA, safety stock is regarded as the only means to deal with supply and demand uncertainty. Therefore, the service time depends on the material availability at the supplying stockpoint. Although a stockpoint might have a deterministic processing time, its entire replenishment time is stochastic due to potential stockouts at upstream stockpoints. The fundamental work by [1] shows that the optimal inventory control rule for such a system is an echelon order-up-to policy. In GSA models, which build on the work by [5], other uncertainty countermeasures besides safety stock exist, so-called operating flexibility, which comprise, e.g., overtime or accelerated production. Therefore, each stockpoint quotes a service time to its successor that it can always satisfy. After this deterministic time span, the successor receives all requested materials and can start processing. While both approaches have been used in inventory theory for a long time, so far no detailed comparison has been drawn between them.

The purpose of this paper is to compare the two modelling approaches described in Sect. 2 and gain general insights into their relative performance by analyzing the simplest version of a supply chain, a two-stage serial one. Section 3 reports the results of a simulation study and points out the main findings. In particular, the service time guarantee of the GSA is explicitly taken into account in the simulation model. Thus, the effect of having a single means to deal with demand uncertainty versus multiple means can be investigated.

2 Multi-Echelon Safety Stock Optimization Approaches

In the SSA and the GSA, each stockpoint i in the supply chain performs a certain processing function, e.g., a step in a manufacturing or transportation process, and is a potential location for holding safety stock after the process has finished. The numbering of stockpoints i is done in increasing order from the most upstream to the most downstream one, $i = 1, 2, \dots, n$. The processing time λ_i is assumed to be deterministic and an integer multiple of a base period. No capacity constraints exist. Customer demand is assumed to be stationary and independent across non-overlapping intervals with mean μ and standard deviation σ . f_λ denotes the λ -period demand probability density. Unsatisfied customer demands are backordered. Without loss of generality, production coefficients are set equal to 1. Moreover, all stockpoints operate a periodic-review base-/echelon-stock policy with a common review period.

2.1 Stochastic-Service Approach

In the SSA, stock insufficiencies at the supplying stockpoint delay the delivery of the stockout quantity until new material becomes available from incoming orders. Consequently, the replenishment time L_i at each stockpoint i , $L_i = ST_{i-1} + \lambda_i$, is stochastic as it consists of the stochastic service time of the predecessor of i , ST_{i-1} , and its own processing time, λ_i . To achieve a predefined external service level, sufficient stock has to be held to cope with demand variability during the replenishment time. Due to the echelon order-up-to policy, each stockpoint takes into account all available amount of stock in the downstream part of the supply chain and orders such that the external service level is achieved. Whereas the external service level is given exogenously, internal service levels are decision variables within the optimization.

Due to the decomposition result derived by [1], optimal echelon order-up-to levels, S_i , can be obtained starting with the final-stage stockpoint (see [6] for details). For the two-stage case, S_i have to be set such that

$$\int_0^{S_2} f_{\lambda_2}(u) du = \frac{p + h_1}{p + h_2} \quad \text{and} \quad \int_0^{S_2} \int_0^{S_1 - u} f_{\lambda_1}(v) f_{\lambda_2}(u) dv du = \frac{p}{p + h_2},$$

where S_i and h_i denote the echelon order-up-to level and holding cost of stockpoint i , respectively. For a given out-of-stock probability $1 - \alpha$, a so-called α -service level, the required penalty cost per item short, p , can be determined by $\alpha = p/(p + h_2)$. The local base-stock level of i is given by the difference of echelon order-up-to levels of adjacent stockpoints, $S_i - S_{i+1}$. An exact mathematical solution for more complex

supply chain structures requires considerable computational effort due to numerical integration.

2.2 Guaranteed-Service Approach

Since the GSA assumes operating flexibility, only internal demand variability up to a maximum reasonable level at a stockpoint has to be covered by safety stocks. Excessive demand is handled by operating flexibility as a second means. Thus, a stockpoint can always meet its service time commitment.

The maximum reasonable level for each stockpoint can be seen as the result of an internal service level requirement. Therefore, safety stock at stockpoint i , SST_i , can be determined by the well-known square root safety stock formula of the single-echelon case, i.e. $SST_i = k_i(\alpha_i) : \sigma : \sqrt{T_i}$ where k_i represents the safety factor which depends on the service level α_i and the lead time demand. T_i denotes the time span for which safety stock has to be held. Due to the coupling of adjacent stockpoints via service times in the multi-echelon case, T_i does not represent the replenishment time L_i , but the net replenishment time, $T_i = ST_{i-1} + \lambda_i - ST_i$, i.e. the replenishment time of i , $ST_{i-1} + \lambda_i$, minus coverage requirements postponed to successors through a (positive) service time, ST_i . The optimization problem for the two-stage serial supply chain is given by

$$\begin{aligned} \min C = & \sum_{i=1}^2 h_i : k_i(\alpha_i) : \sigma : \sqrt{ST_{i-1} + \lambda_i - ST_i} \\ \text{s.t.} \quad & 0 \leq ST_1 \leq ST_0 + \lambda_1 \quad , \quad ST_0 = ST_2 = 0 \quad . \end{aligned}$$

The objective function minimizes safety stock holding costs. Costs for using operating flexibility are not included explicitly, but only implicitly by setting the internal service levels α_i . The constraints ensure that the service time of a stockpoint must not exceed the replenishment time and not be negative. For the external supplier and the final stockpoint a service time of 0 is assumed, but not necessarily. In an optimal solution, the service time of a stockpoint is either 0 or equals its replenishment time (see [5]). The global optimum across all stockpoints can be found by dynamic programming (see [2]), even for complex supply chain structures (see [4]).

Note that in contrast to the SSA, which makes its cost calculations based on an exact assessment of stocks in the supply chain, the minimization of safety stocks in the GSA is only an approximation since safety stock represents the average net stock and therefore includes negative values in a backordering situation.

3 Simulation Study

In order to be able to include and assign costs for operating flexibility, where no simple analytic expressions exist, a simulation model is used. The study has been conducted in ARENA v7.01 for a two-stage serial supply chain with the following parameters: $\lambda_1, \lambda_2 \in \{1, 5\}$ including the review period at the final stage; $h_1 = 10$, $h_2 \in \{15, 20, 25\}$; normally i.i.d. demand with $\mu = 100$ and $\sigma \in \{20, 40\}$; external α -service level $\in \{0.8, 0.85, 0.9, 0.95, 0.99\}$; operating flexibility cost $c_1 \in \{20, 40\}$. In total, 240 scenarios have been analyzed. For each scenario, 10 simulation runs

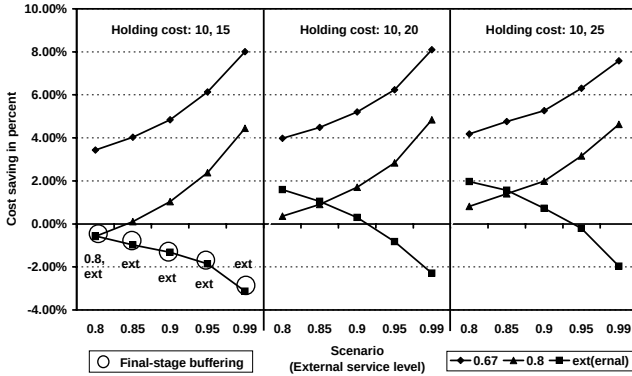


Fig. 1. Cost Advantage of GSA over SSA ($\lambda_1 = 5$, $\lambda_2 = 1$, $\sigma = 20$)

with 20,000 periods each have been conducted. Optimal base-stock levels have been calculated in MAPLE v10. In case of the SSA, the algorithm described in [6] has been used. The incomplete convolution for the normal distribution has been determined using numerical integration. For the GSA, results have been obtained by a dynamic program. In the simulation model, operating flexibility is implemented such that, in case of a stockout, the stockout quantity at a stockpoint is speeded up either from its pipeline stock or from physical and pipeline stock of predecessors at the beginning of a period.

In a first step, internal service levels in the GSA are set equal to the external service level α . This approach is implemented in many relevant papers without justification or simply by stating that this is done in accordance with management (see, e.g., [3]). Interestingly, it results in higher costs for the GSA in the majority of the investigated scenarios with an average cost inferiority of about 1%. Only in very few cases where α is below 90%, the GSA causes slightly lower costs (see Fig. 1), although an operating flexibility cost of only 20 is assigned. The reason is that safety stocks at the upstream stockpoint are sized with respect to the high external service level. That is why, in many scenarios the GSA advises safety stocks to be held solely at the final stockpoint (final-stage buffering), because the reduction in the safety stock amount due to the square root effect at the downstream stockpoint dominates the holding cost advantage at the upstream one. Only for high downstream holding cost, e.g. $h_2/h_1 = 150\%$, safety stocks are carried at both locations. In the SSA, optimal internal service levels range below 80% in 188 of the 240 scenarios and thus more stocks are held upstream. Therefore, the internal service level represents a major driver for the safety stock size and allocation and a more cost-oriented determination of the internal service levels is recommended.

Since the operating flexibility cost c_i has to be incurred for every shortfall item, it can be regarded as the penalty cost in the single-echelon newsvendor formula. Consequently, a pragmatic approximation for the required internal service level is $\alpha_i = c_i / (c_i + h_i)$ (see [4], p.36), resulting in internal service levels of 66.7% and 80% for $c_i = 20$ and 40, respectively. Using these cost-based internal service levels causes major cost reductions compared to the “internal = external” service level case, and in 192 of the 240 scenarios even leads to a cost superiority of the GSA over

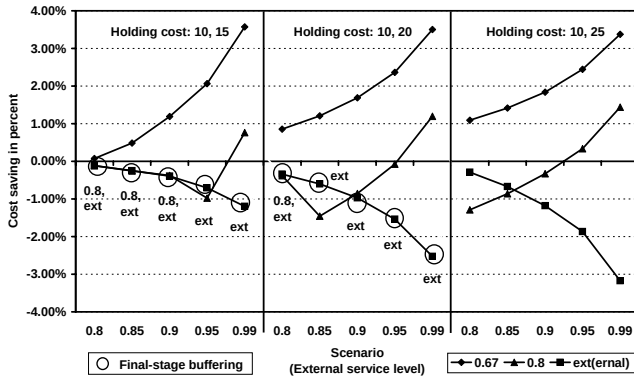


Fig. 2. Cost Advantage of GSA over SSA ($\lambda_1 = 5$, $\lambda_2 = 5$, $\sigma = 40$)

the SSA (see Figs. 1, 2). Obviously, the cost advantage decreases with growing cost for operating flexibility. However, even if the operating flexibility cost is four times the holding cost of the stockpoint, the maximum cost disadvantage of the GSA is only about 1.5%. Cost savings, on the other hand, are much higher in most cases, amounting to about 12% at maximum. In cases where the GSA prescribes safety stock at both locations, the cost advantage increases monotonically, the higher the external service level is set. In order to achieve a higher external service level, the SSA increases the stock level at both locations, in particular at the upstream one to reduce possible stockouts. In the GSA, the stock level of the upstream stockpoint remains constant and only the downstream one is increased. Potential stockout situations are counteracted by operating flexibility. In situations where the internal service level equals the external one, the reverse can be observed. The GSA performs worse, the higher the external service level gets due to the large internal stock requirement.

Processing time characteristics have an influence in so far as a larger cost benefit can be gained from the GSA when a long processing time is followed by a shorter (or equal) one. The safety stock required for the longer processing time at the upstream location is sized with regard to the cost-based service level as opposed to the downstream stockpoint where the high external service level has to be met. This cost effect holds for both coefficients of demand variation. Demand variability affects the slope of the cost advantage curve which gets steeper for a larger coefficients of variation. For low external service levels, the cost advantage decreases whereas it increases for high service levels.

4 Conclusion

The simulation results show that companies can benefit from having multiple means available to cope with demand uncertainty as assumed by the GSA. The right specification of the internal service levels is crucial, however. Simply setting internal service levels equal to the external one without taking into account the operating flexibility cost leads to a cost inferiority of the GSA for high external service levels and, therefore, favors a pure safety stock strategy as assumed by the SSA. From a

practical point of view, the computational attractiveness of the GSA also backs its application. However, if operating flexibility is not available at certain stockpoints of the supply chain, the SSA is the only possible one to use.

Based on this work, several issues are worth to be addressed. In practice, the use of fill rate service measures instead of a stockout occurrence related one is common. The extension to this type of service measure as well as the inclusion of other supply chain structures, such as divergent or convergent systems, are of primary interest. Moreover, a “hybrid-approach” which optimizes the service approach type to use at each stockpoint in a supply chain might be more realistic than exclusively using one approach for all stockpoints.

References

1. Clark AJ, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. *Management Science* 6:475–490
2. Graves SC (1988) Safety stocks in manufacturing systems. *Journal of Manufacturing and Operations Management* 1:67–101
3. Graves SC, Willems SP (2003) Supply chain design: Safety stock placement and supply chain configuration. In: de Kok AG, Graves SC (eds) *Supply chain management: Design, coordination and operation*. North-Holland, Amsterdam
4. Minner S (2000) *Strategic safety stocks in supply chains*. Springer, Berlin Heidelberg New York
5. Simpson KF (1958) In-process inventories. *Operations Research* 6:863–873
6. van Houtum GJ, Inderfurth K, Zijm WHM (1996) Materials coordination in stochastic multi-echelon systems. *European Journal of Operational Research* 95:1–23

A Stochastic Lot-Sizing and Scheduling Model

Sven Grothklags¹, Ulf Lorenz¹, and Jan Wesemann²

¹ Institut für Informatik, Paderborn University, Germany

² Wirtschaftsinformatik, insb. CIM, Paderborn University, Germany

1 Introduction

In the academic world, as well as in Industry, there is a wide unity that companies with a combination of high customer-value products and good logistics are especially successful in competition. Good logistics are marked by low production- and inventory costs as well as by a high ability to deliver timely.

Although the benefit of good logistics is indisputable, there are quite different ideas, how to achieve the goal. One approach begins with the formulation of a mathematical optimization model, which monolithically describes the planning system. The model allows to determine and to compare different admissible solutions. Moreover, various solution procedures and algorithms can be evaluated in a systematical way. A weakness of the currently published models is that input data are typically assumed to be deterministic. Approaches which combine dynamic capacitate planning with considering stochastic influences are still missing [13]. A different, however, very important approach with its emphasis more on controlling than on planning, mainly used for stock production can e.g. found in [8].

We present a new model, based on a direction as proposed e.g. in [13, 6, 2]. Injecting a certain amount of stochasticity into the data of such model, leads us to the fields of Decision Making Under Uncertainty [11], On line Algorithms [5], Stochastic Optimization [4] etc. Recently, we saw that stochastic information can indeed be exploited [1, 3, 7, 9, 12]. All these stochastic approaches fall into a category which Papadimitriou called 'Games against Nature' [10]. Interestingly, the typical complexity class of these games is PSPACE, the same complexity class to that many two-person zero-sum games belong.

In the next section, we elaborate such a model for production planing, and thereafter, we give a short outlook.

2 Production Planning as a Game Against Nature

2.1 The Deterministic Small-Bucket Multi-Level Capacitated Lot-Sizing and Scheduling with Parallel Machines (D-SMLCLS-PM)

We are interested in a combined problem of lot sizing and scheduling with the following properties: Several items have to be produced to meet some known dynamic demand. A general gozinto-network provides the 'is-piece-of'-relations between the items. Backlogging, stockouts, and positive lead times are considered.

The production of an item requires the exclusive access to a machine, and items compete for these machines. A specific item may be produced alternatively on different machines and can have different production speeds on single machines. Even the production speed on the same machine can vary over time.

The production of an item on a machine can only take place if the machine is in a proper state. Changing the setup state of a machine from one state to another causes setup costs and setup times, which both may depend on the old and new state. Setup states are kept until another setup changes occur.

The planning horizon is divided into a finite number of time periods. Items which are produced in a period to meet some future demand must be stored in an inventory and cause item-specific holding costs.

One restriction which we invented to the model is that setup changes are only allowed to start at the beginning of a period. As a consequence, every machine can produces at most one item type per period.

The following variables are used to encode a solution:

Symbol	Definition
$q_{i,t}$	Production quantity of item P_i by task i in period t
$s_{p,t}$	Inventory for item p at the end of period t
$b_{p,t}$	Backlog of item p at the end of period t
$o_{p,t}$	Shortfall quantity of item p in period t (no backlog)
$x_{i,j,t}$	Binary variable which indicates whether a setup change from task i to j occurs in (at the begin of) period t ($x_{i,j,t} = 1$) or not ($x_{i,j,t} = 0$)
$y_{i,t}$	Binary variable which indicates whether machine M_i is set up for task i at the end of (during) period t ($y_{i,t} = 1$) or not ($y_{i,t} = 0$)

To describe the optimization model we also need the following input parameters:

Symbol	Definition
P	Set of items
M	Set of machines
I	Set of tasks
T	Number of periods ($1 \dots T$)
M_i	Machine needed by task i
P_i	Item produced by task i
I_i	$= \{j \in I : M_j = M_i, j \neq i\}$; Set of tasks using the same machine as task i

S_p	Nonnegative holding cost for having one unit of item p one period in inventory
B_p	Nonnegative cost for backlogging one unit of item p for one period
O_p	Stockout cost for not fulfilling the (external) demand of one unit of item p
$X_{i,j}$	Nonnegative setup cost for switching from task i to j (on machine $M_i = M_j$)
Q_i	Production cost for one unit of item P_i with task i
$D_{p,t}$	External demand for item p in period t
$A_{p,q}$	Gozinto-factor. It is zero if item q is not an immediate successor of item p . Otherwise: the quantity of item p that is directly needed to produce one item q
$\Delta_{p,i}$	(Integral) number of periods for transporting item p to machine M_i (for task i)
$C_{i,t}$	(Maximum) production quantity of item P_i with task i during period t
$T_{i,j}$	(Fractional) number of periods for switching from task i to j
$R_{i,j}$	$= \lfloor T_{i,j} \rfloor$; Number of periods exclusively used for switching from task i to j
$R_{i,j}^f$	$= T_{i,j} - R_{i,j}$; Fraction of the the $(R_{i,j} + 1)$ th period after switching from task i to j used for switching

Now, the following MIP defines our lot sizing and scheduling problem.

$$\text{minimize } \sum_{t=1}^T \left(\sum_{p \in P} (S_p s_{p,t} + B_p b_{p,t} + O_p o_{p,t}) + \sum_{\substack{i \in I \\ j \in I_i}} X_{i,j} x_{i,j,t} + \sum_{i \in I} Q_i q_{i,t} \right) \quad (1)$$

$$\text{subject to } s_{p,t-1} + \sum_{i: P_i=p} q_{i,t} - D_{p,t} - \sum_{\substack{q \in P \\ j: P_j=q}} A_{p,q} q_{j,t+\Delta_{p,j}} + b_{p,t} + o_{p,t} - b_{p,t-1} = s_{p,t} \quad \forall p \in P, \quad t = 1 \dots T \quad (2)$$

$$b_{p,t} + o_{p,t} - b_{p,t-1} \leq D_{p,t} \quad \forall p \in P, \quad t = 1 \dots T \quad (3)$$

$$\sum_{i: M_i=m} y_{i,t} = 1 \quad \forall m \in M, \quad t = 1 \dots T \quad (4)$$

$$y_{i,t-1} + y_{j,t} - 1 \leq x_{i,j,t} \quad \forall i \in I, j \in I_i, \quad t = 1 \dots T \quad (5)$$

$$C_{i,t} \left(y_{i,t} - \sum_{\substack{j \in I_i \\ t-R_{j,i} < t' \leq t}} x_{j,i,t'} - \sum_{j \in I_i} R_{j,i}^f x_{j,i,t-R_{j,i}} \right) \geq q_{i,t} \quad \forall i \in I, \quad t = 1 \dots T \quad (6)$$

$$q_{i,t} \geq 0; \quad s_{p,t}, b_{p,t}, o_{p,t} \geq 0 \quad \forall i \in I, \forall p \in P, \quad t = 1 \dots T \quad (7)$$

$$x_{i,j,t} \in \{0,1\}; \quad y_{i,t} \in \{0,1\} \quad \forall i \in I, j \in I_i, \quad t = 1 \dots T \quad (8)$$

Objective (1) minimizes the sum of the total holding, backlog, stockout, setup, and production costs.

Equations (2) are the inventory balances. The inventory of item p at the end of period t can be expressed by the inventory at the end of period $t - 1$ plus what was produced during period t minus external and internal demand. Furthermore, we must add the amount of items that are backlogged and canceled in period t (these items reduce the external demand in the current period) and subtract the number of items backlogged in period $t - 1$ (these items must still be produced - now or in some future period). Note that for the internal demand, positive lead times $\Delta_{p,j}$ are taken into account by taking item p out of the inventory $\Delta_{p,j}$ periods before starting the demanding task j . No inventory costs for the internal demand of item p occur during these $\Delta_{p,j}$ periods but these costs can be added to the production costs of task j , if desired.

Inequalities (3) ensure that the number of items (additionally) backlogged and canceled does not exceed the external demand during each period. Therefore, the internal demand can always be fulfilled and production is guaranteed.

Equations (4) together with (8) ensure that every machine has exactly one setup state at the end of every period.

Inequalities (5) together with (8) identify setup changes. If a machine was in state i at the end of period $t - 1$ ($y_{i,t-1} = 1$) and the state at the end of period t is j ($y_{j,t} = 1$) a setup change from i to j occurred during period t and inequality (5) forces $x_{i,j,t}$ to one. For all other setup states $x_{i,j,t}$ can become zero and positive setup costs $X_{i,j}$ will enforce this for optimal solutions. Note that therefore conditions (8) can be replaced by $x_{i,j,t} \geq 0$ without changing the set of optimal solutions.

Production capacity constraints are expressed by inequalities (6). First of all, a task i can only produce items in period t if machine M_i is in the proper setup state during this period ($y_{i,t} = 1$). If all occurring $x_{i,j,t}$ variables are zero we can then produce up to $C_{i,t}$ items in period t . As we want to take setup times into account in our model, we must also make sure that the setup change from some task j to i has been finished before (or during) period t . The first sum effectively reduces the available production capacity to zero for periods which are fully reserved for setup changes. The second sum cuts of the necessary fraction of the production capacity for periods during which a setup change finishes. Note that because of (7) the left hand side of an inequality of type (6) must not become negative. Thus all $x_{i,j,t}$ variables must be zero if $y_{i,t}$ is zero and at most one $x_{i,j,t}$ variable can be one if $y_{i,t}$ is one. This ensures that a machine remains in its setup state i during the whole setup change leading to state i and that no two setup changes can overlap.

The constraints (7) are the non-negativity conditions for production quantity, inventory, backlogging, and stockout variables. Constraints (8) define our binary-valued setup change and setup state variables.

With a bit more effort, the model can be extended in such a way that integral lot sizes are modeled as well.

2.2 The Stochastic Small-Bucket Multi-Level Capacitated Lot-Sizing and Scheduling with Parallel Machines (S-SMLCLS-PM)

On the basis of the previous section, we can now move from the D-SMLCLS-PM to a stochastic version, the S-SMLCLS-PM. In contrast to the deterministic version, we now assume that customer demands are not described with a natural number, but with the help of a finite probability distribution, which reflects that customer demands are not exactly known at planning time. Moreover, the production speed of the machines is described with the help of a probability distribution, as well.

We formalize the stochastic planning problem as a game. Before we go into details, let us inspect the notion of time. Let T_0 be the start time, T_e the end time of our game, T_i any time point between T_0 and T_e . Similar as in the deterministic case, the time is divided into intervals, into small time periods. Let the timespan between T_i and T_{i+1} be δ . We assume that everything that takes place between time T_i and $T_i + \delta$ happens simultaneously, at time point $T_i + \delta$. Let τ_{det} be a time period, such that we can assume demands with deadlines between T_i and $T_i + \tau_{det}$ to be deterministic. τ_{det} is a multiple of δ . Summarizing, we assume two different time blocks: deterministic knowledge about future external demands shortly behind the here-and-now time-line, as well as the sizes of past lots. Moreover, we assume to know finite distributions about demand-data beyond $T_i + \tau_{det}$, and finite distributions about future production capacities per period, i.e. the $C_{i,t}$.

The game has two players who move in an alternating way. One player is a company that tries to maximize its profits. The other player is Nature, who realizes probability distributions, rolls a die, respectively.

The actions of Nature can be described as follows. At time T_i , Nature determines the size of those lots which have been finished between T_i and T_{i-1} . Moreover, it determines those demands, whose deadlines are between T_i and $T_i + \tau_{det}$.

The company has to assign tasks to the machines (= provide a plan) at least from time T_i to T_{i+1} at each time point T_i . The aim is to maximize the profit over the full period T_0 to T_e . Thus, instead to provide one plan, the company provides a sequence of plans $P_0, \dots, P_e$ from time T_0 to time T_e . Obviously, the optimization task is a multi-stage decision process. Gains and costs of a realized run are gathered concerning Equation 1. Here, we see the relation between the stochastic and the deterministic optimization problems. From a posteriori point of view, the originally fuzzy input data have been determined and each run has to follow the production rules of deterministic problem description. Moreover, the costs and profits are determined with the help of the deterministic model. Note, however, that because of imperfect information at decision time, an optimal decision at a time point T_i is not necessarily identical with the optimal decision under the assumption that data realize as they do. An optimal decision at a time point T_i can only be optimal in expectation.

As soon as Nature has made a move, the company has to re-act. Ideally, the company instantly reacts.

3 Conclusion and Future Work

We have introduced a new production model which is mainly geared to the deterministic lotsizing and scheduling models for deterministic planning. Stochastic influences are additionally injected and as a result, we described the optimization problem in form of a Game against Nature.

The S-SMLCLS-PM-model allows to compare algorithms experimentally in a clearly defined optimization world. E.g., we can keep a certain, especially interesting problem instance fixed, and then make a lot of simulation runs. Because the input data are probability distributions with nice properties, we can conclude to the quality of an algorithm via statistics. Another way to evaluate an algorithm is to arrange competitions as we know them from game playing. We are going to implement a simulation world for the S-SMLCLS-PM-model, which allows to compare different algorithms with the help of statistics and with the help of competitions. We aim at realistic test instances like a complete motor of a car.

References

1. R. Bent and P. van Hentenryck. Scenario based planning for partially dynamic vehicle routing problems with stochastic customers. *Operations Research*, 52(6), 2004.
2. A. Drexel and K. Haase. Proportional lot sizing and scheduling. *International Journal of Production Economics*, 40:73–87, 1995.
3. J. Ehrhoff, S. Grothklags, and U. Lorenz. Parallelism for Perturbation Management and Robust Plans. *Proceedings of the 11th International Euro-Par Conference*, (J.C. Cunha, P.D. Medeiros eds), Springer LNCS 3648, pages 1265–1274, 2005.
4. S. Engell, A. Märkert, G. Sand, and R. Schultz. Production planning in a multiproduct batch plant under uncertainty. *Preprint 495-2001, FB Mathematik, Gerhard-Mercator-Universität Duisburg*, 2001.
5. A. Fiat and G. Woeginger. Online Algorithms: The State of the Art. *LNCS*, 1998.
6. A. Kimms and A. Drexel. Proportional lot sizing and scheduling: Some extensions. *Networks*, 32:85–101, 1998.
7. A. Kleywegt and A. Shapiro. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 15(1-2), 1999.
8. H. Löffding. Verfahren der Fertigungssteuerung. *Springer, Berlin*, 2004.
9. R. Moehring, A. Schulz, and M. Uetz. Approximation in stochastic scheduling: the power of lp-based priority policies. *Journal of ACM*, 46(6):924–942, 1999.
10. C. H. Papadimitriou. Games against nature. *Journal of Computer and System Science*, 31:288–301, 1985.
11. A. Scholl. Robuste Planung und Optimierung. *Physiker Verlag*, 2001.
12. D. Shmoys and C. Swamy. Stochastic Optimization is (almost) as easy as Deterministic Computation. *Proceedings of Foundations in Computer Science (FOCS-2004)*, pages 228–237, 2004.
13. H. Tempelmeier. Material-Logistik, Modelle und Algorithmen für die Produktionsplanung und -steuerung in Advanced Planning-Systemen. *Springer, 6. Aufl., Berlin*, 2006.

A Disassemble-to-Order Heuristic for Use with Constrained Disassembly Capacities

Tobias Schulz

Faculty of Economics and Management, Otto-von-Guericke University Magdeburg,
Germany
`tobias.schulz@ww.uni-magdeburg.de`

1 Introduction

Due to the increasing environmental awareness of their customers, manufacturing firms have to incorporate an ecological aspect in their corporate identity to remain competitive. In several cases these firms offer an option for their customers to return the product to them after its end-of-use or alternatively its end-of-life. The field of reverse logistics presents an overall approach on how these backward flows of products can be handled efficiently. Therefore, it expands the original focus of logistics to a closed-loop supply chain perspective.

Remanufacturing is one of the manifold options to deal with the returned products [2]. Within this process the returned products (which we will refer to as cores) are disassembled to obtain functional components (which we will refer to as leaves) that can be reassembled into “good-as-new” products afterwards. Before reassembly these components must be cleaned, inspected and if necessary reworked to ensure full functionality. A part of the embedded economic value of those reused “good-as-new” leaves can thus be saved with this procedure. For those components that cannot be reused, newly produced parts have to be availed. The remanufacturing of engines gives a practical example for this process where “good-as-new” engines are offered to the customers in exchange for broken ones. Since reworked components are used, both the environment because of reduced landfilling and the customers because of cheaper spare parts are satisfied.

A known demand for the remanufactured products leads by a bill-of-materials explosion to an individual demand for each leaf. This demand can either be fulfilled by procuring the leaves or by disassembling a specific amount of cores. The so called disassemble-to-order problem seeks to find a solution for this challenging planning task. Since an optimal solution for realistic problem sizes can hardly be found within a reasonable time, as can be seen in section 2, the use of heuristics can be motivated. Prior work on heuristics for the disassemble-to-order problem (see e.g. [1]) has neglected restricted disassembly capacities. However, our experience with several engine remanufacturers has shown that an unrestricted disassembly capacity does not portray reality. As not every desired core can be disassembled with restricted capacities, a more expensive planning results because several leaves

have to be procured instead of being obtained through the usually cheaper option of disassembly.

Following this short introduction, an integer programming model to solve this problem to optimality is presented. After that, a heuristic approach that can be applied to the disassemble-to-order problem with restricted disassembly capacities is introduced in section 3 and a numerical example is shown in section 4. Finally, a short conclusion and a brief outlook is given in the last section.

2 Exact Solution Method

We will now put forth an integer programming model which can be used to find the optimal solution to a disassemble-to-order problem. We can start by examining the underlying assumptions. All the data that is put into the model is of deterministic nature, which means that the demand, the yield rates (amount of leaves that is obtained through the disassembly of a core), the capacity absorption and the overall capacity are known with certainty. With respect to the capacity absorption factors it should be mentioned, that these are modelled as linear factors since the often manual process of disassembly is not significantly influenced by setup times. Further assumptions are that there is no disassembly lead time considered and that the number of cores that can be bought from the market is not constrained. The last assumption that should be mentioned here is that cores are completely disassembled down to their leaves.

A disassemble-to-order problem can be solved to optimality with the aid of an integer programming model. This section is dedicated to introduce all required variables of the problem setting. Starting with the indices, i represents a single core and I represents the set of considered cores. The index j represents a single leaf while J is the set of leaves. There is also the time index t , which indicates that a multiperiod problem of T periods is examined. The capacity absorption factors a_i represent the amount of capacity that is needed for disassembling a core i . The overall disassembly capacity in each period is called C_t . While the demand for each leaf in period t is described by D_{jt} , the deterministic yield rates are represented by n_{ij} . There are several cost parameters in this model, namely c_i^s which is the overall core cost (containing the costs of procuring, transporting and disassembling a core). c_j^p represents the cost for procuring a leaf j , while its disposal costs c_j^d and holding it for one period results in costs of c_j^h . Solving the integer program results in deciding on the amount of core i to be disassembled in period t , x_{it}^s , as well as the amount of leaf j to be procured and the amount of this leaf to be disposed of in period t , x_{jt}^p and x_{jt}^d , respectively, and the height of the inventory level at the end of period t , named y_{jt} . The starting inventory is denoted y_{j0} . The integer programming model is given as:

$$\text{Min } Z = \sum_{t=1}^T \left[\sum_{i \in I} c_i^s \cdot x_{it}^s + \sum_{j \in J} (c_j^p \cdot x_{jt}^p + c_j^d \cdot x_{jt}^d + c_j^h \cdot y_{jt}) \right] \quad (1)$$

$$y_{jt} = y_{j,t-1} + \sum_{i \in I} x_{it}^s \cdot n_{ij} + x_{jt}^p - x_{jt}^d - D_{jt} \quad j \in J \quad t = 1, \dots, T \quad (2)$$

$$y_{jT} = 0 \quad j \in J \quad (3)$$

$$\sum_{i \in I} a_i \cdot x_{it}^s \leq C_t \quad t = 1, \dots, T \quad (4)$$

$$x_{it}^s, x_{jt}^p, x_{jt}^d, y_{jt} \geq 0 \text{ and integer } \forall i, j \quad t = 1, \dots, T$$

The objective function (1) tries to minimize the sum of all relevant costs (core costs as well as leaf procurement, disposal and holding costs). Equation (2) represents the inventory balance constraint for every leaf in each period. It declares that the inventory consists of the inventory of the last period plus its inflow of leaves (external procurement, yield from disassembly) and minus its outflows (demand, disposal). Equation (3) ensures that no leaf is held at the end of the planning horizon. The compliance of the capacity restriction in each period is ensured by equation (4). Finally, non-negativity and integrality constraints are given. Such an integer programming model can be solved to optimality with the aid of commercial software tools (e.g. XPRESS-MP). Due to the fact that for realistic problem sizes (many cores, many leaves) the number of integer variables is very large, the optimal solution can not often be found within a reasonable amount of time which motivates the use of a heuristic solution method.

3 Heuristic Approach

The following heuristic approach is applicable to the problem formulation described above. Within two distinct phases the heuristic tries to find a good solution. At the beginning of the first phase the whole demand for leaves is entirely satisfied by procuring them individually. The heuristic iteratively improves this non-disassembly solution by finding the most valuable core to be disassembled in each iteration. Both phases of the heuristic can be described by the same following pseudocode (slight differences occur in the steps 2 and 3). After reaching the end of the first phase, the second phase uses the solution obtained by the first phase and starts at step 1 of the pseudocode. The heuristic's final solution follows the end of phase two in the final period.

1. **start** $t = 1$
2. calculate / update net requirements
3. determine the core with the highest overall saving value
4. **if** current capacity is not violated **and** highest overall saving value > 0
 - a) **then** disassemble one unit of the core with the highest overall saving value
 and goto 2.
5. **end if**
6. **if** $t + 1 \leq T$ **then**
 - a) $t = t + 1$ **and goto** 2.
7. **end if**
8. **end**

Each iteration of the first phase begins with the calculation of the leaves' net requirements of the current period t (step 2). We refer to the term net requirement as the unsatisfied part of the demand for leaves in the current period since previous iterations have made disassembly decisions that already satisfy a part of the current demand. The heuristic algorithm proceeds by determining the overall saving for each core individually (step 3). The first phase of the heuristic neglects any holding and disposal decision, as it only regards the net requirements of the current period. To incorporate those decisions, one would have to consider possibly existing demands of future periods. While this is obvious for holding decisions, the idea behind forbidding disposal decisions is that any superflously obtained leaf in the current period could be held in inventory to satisfy a future demand instead of disposing it of. Therefore, the overall saving of a core in phase one results from its profitability (value of all leaves contained therein for which there is a demand minus the overall core costs) multiplied with the maximum amount of that core type that can be disassembled without disposing of a single leaf. This overall saving value of a core represents the absolute improvement of the current solution if there is only the option to disassemble this specific core type. Of course, this can only be taken as a signal for the disassembly decision to be made, since all other cores are disregarded. If the capacity constraint is fulfilled and the highest overall saving of all cores is positive, one unit of that core is disassembled and the next iteration starts. If this is not the case, the next period is considered until the end of the planning horizon is reached and phase two starts.

After finishing the first phase, the iterative solution improvement of the second phase begins with the calculation of the net requirements. In contrast to the first phase not only the net requirements of the current period t , but all net requirements until the end of the planning horizon T are updated. That is because the calculation of each core's overall saving value in the second phase explicitly regards future demands. As a starting point the disassembly of one additional unit of a core is considered. Due to the fact that there is at least one leaf for which there is no net requirement in the current period, additional costs are incurred by the disassembly of this core. The heuristic has to check therefore whether there is a future demand for these leaves. If this is the case, the considered leaves will be held in inventory. If not, they will be disposed of. Therefore, the profitability criterion of phase one has to be adopted to incorporate the cost of disposal and/or holding. If this adopted profitability is positive, the profitability of one further disassembled unit is calculated. The overall saving of that core type is determined by adding up all marginal profitabilities as long as they are positive and the capacity constraint is fulfilled.

As in phase one, each iteration ends with the disassembly of one unit of the core that signals the highest overall saving if all constraints (capacity, positive saving) are fulfilled.

This heuristic approach can not only be applied to the case of constrained disassembly capacities but also to the uncapacitated case. Step 4 of the pseudocode is the only calculation step that has to be changed by disregarding the capacity conditional in order to implement the introduced heuristic to the case of unrestricted disassembly capacities.

4 Numerical Example

To illustrate the heuristic algorithm a simple numerical example has been chosen. This example is tested on the one hand with a limited disassembly capacity of four cores per period, regardless their type, and on the other hand with an unlimited capacity. Two different cores (numbered 1 and 2) are regarded in this example which obtain through disassembly four different leaves (numbered 3 to 6). The core costs are given as 10 for core 1 and 14 for core 2, respectively. Table 1 shows the other relevant parameters for this example.

Table 1. Data for numerical example

	Leaf 3	Leaf 4	Leaf 5	Leaf 6
Yield of core 1	0	3	1	1
Yield of core 2	2	1	0	1
Demand in period 1	7	9	13	1
Demand in period 2	4	6	12	1
Demand in period 3	6	9	16	4
Demand in period 4	10	5	10	1
Leaf disposal cost	2.5	0.7	1.8	0.5
Leaf holding cost	0.08	0.04	0.03	0.07
Leaf procurement cost	7.7	4.3	3.4	7.0

Table 2 shows the core disassembly decisions in each of the four periods that were computed by applying the introduced heuristic approach to the problem. As one can see, only a fracture of the desired cores that could be disassembled with unlimited capacity can be disassembled in the capacitated scenario. It should be mentioned here that this kind of behavior cannot generally be observed in every example. For the sake of brevity, we omit the presentation of the other decision variables.

The optimal solutions were computed with the integer programming model described in section 2 to compare it with the heuristic solution of the example. At the unconstrained disassembly capacity scenario the heuristic solution shows a relative cost deviation of 0.71 % to the optimal solution. In the constrained capacity case the heuristic solution deviates by only 0.29 % from the optimal one.

Table 2. Heuristic core disassembly decisions

i / t	capacitated scenario				uncapacitated scenario			
	1	2	3	4	1	2	3	4
1	1	1	2	0	2	1	2	0
2	3	3	2	4	10	0	2	1

5 Conclusion

As the required solution time increases significantly for a disassemble-to-order problem applied to practical cases, the use of heuristic approaches is conceivable. The introduced heuristic in this paper has shown for a small example that it is able to handle this problem with constrained capacities. Of course, a performance study has to be conducted to examine the heuristic’s performance thoroughly under various parameter settings. This study would also reveal the worthwhile information, on how the performance of the heuristic is affected by the tightness of capacity.

References

1. I.M. Langella and T. Schulz (2006) Effects of a rolling schedule environment on the performance of disassemble-to-order heuristics. In: *Proceedings of the Fourteenth International Working Seminar on Production Economics*, Vol. 1, pages 241-256. Innsbruck, Austria.

2. M. Thierry, M. Salomon, J. Van Nunen and L.N. Van Wassenhove (1995) Strategic issues in product recovery management. *California Management Review*, 37(2):114-135.

Supply Chain Management and Advanced Planning in the Process Industries

Norbert Trautmann and Cord-Ulrich Fündeling

Departement Betriebswirtschaftslehre, Universität Bern, Switzerland

`trautmann@ifm.unibe.ch`

`fündeling@ifm.unibe.ch`

Summary. Advanced Planning Systems decompose Supply Chain Management into long-term, mid-term, and short-term planning tasks. In this paper, we outline these tasks and reference application-oriented state-of-the-art solution approaches with regard to the process industries.

1 Overview

In the recent past, Advanced Planning Systems (APS) have been developed to support Supply Chain Management (SCM). For each section of the supply chain, i.e., procurement, production, distribution, and sales, APS offer support for long-term, mid-term, and short-term planning while always taking into account limited availability of resources. Planning is supported by appropriate demand forecasts (cf. [11]). Long-term planning concerns the product program, the decoupling points, the cooperations with suppliers, the plant locations, the production systems, and the distribution structure. Mid-term planning synchronizes the material flows along the supply chain. Short-term planning is performed for each section of the supply chain individually and concerns order quantities, production schedules, warehouse replenishment, as well as order acceptance and fulfillment. For a general definition of these planning tasks see [14].

In this paper, we examine the planning tasks with regard to the process industries (e.g. chemical, pharmaceutical, food and beverages, or metal casting industries). In those industries, value is added to materials through the successive execution of chemical or physical transformation processes such as heating, filtration, or mixing (cf. [3]). Important peculiarities of process industries that have to be taken into account by SCM are discussed in [1, 16, 23, 35]. In the following, we concentrate on the related planning tasks, i.e., choice of product program, network design, selection of production mode, design of plant layout, mid-term planning, and short-term production planning.

In the remainder of this paper, we outline these planning tasks and briefly review state-of-the-art methods for long-term (cf. Section 2), mid-term (cf. Section 3), and short-term (cf. Section 4) SCM in the process industries. Comprehensive literature

surveys about some of those planning tasks are provided by [23, 25, 32, 33]. A review addressing rework can be found in [13].

2 Long-Term Planning

The planning horizon considered in long-term planning typically comprises three to ten years. An important decision to be made is the choice of the **product program**. In the process industries, each new product generally has to pass a series of extensive regulatory tests. Product program decisions have to take into account the uncertainty that one of these tests fails which prohibits the product from entering the market. The planning problem therefore consists in choosing a set of candidate products and a schedule for their development such that constraints on the development sequence and on the scarcity of research resources are met and the expected net present value is maximized.

Many solution approaches are based on mixed-integer linear programming. A review is contained in [27]. These approaches mainly differ in the way the research resources are considered. In [31], one such model is applied to an industrial case study. Other approaches make use of simulation models. [4] simulate the evolution over time of the new product development process in order to assess product candidates. In the genetic algorithm presented by [5] product development sequences for a set of candidate products are evaluated by simulation.

Network design deals with the determination of plant locations and the distribution structure. A review of solution approaches can be found in [37]. Real-world applications are discussed in [24] and [38].

Furthermore, continuous, batch, or some hybrid **production mode** has to be chosen (cf. also Section 4): In continuous production mode, material flows are continuous. Batch production mode is characterized by discontinuous material flows. In practice, the choice of the production mode is either predetermined or strongly influenced by the types and volumes of products to be produced. For instance, batch production mode is typical for the production of small volumes of specialty chemicals.

Eventually, the **plant layout** has to be established. The problem is to determine the location and orientation of production equipment items (i.e., processing units and storage facilities) such that total construction and operational cost is minimized. Production equipment items may not pairwise overlap, and safety distances between them have to be observed. For a review of solution approaches we refer to [21]. Sometimes, the number and size of or the routing of pipes between the production equipment items have to be determined as well (cf. [12, 21]).

3 Mid-Term Planning

The goal of **mid-term planning** is an efficient utilization of the production, storage, and transportation facilities determined by the long-term planning. The planning horizon has to cover at least one seasonal cycle to be able to balance all demand peaks and is divided into periods with a length of a week up to a month. Decisions have to be made on purchasing, production, and transportation quantities for final

products and intermediates as well as manpower overtime. Typically, the objective is either to meet demand at minimal total purchasing, production, transportation, and holding cost, or to maximize profit. In the latter case, demand must not necessarily be fulfilled. Constraints vital to the process industries arise from the scarcity of production and storage resources, quarantine or shelf life times for intermediates or final products, and general (including cyclic) material flows.

Most solution approaches start from a mixed-integer linear programming formulation. Due to the size and complexity of the problem, standard software often fails even to find a feasible solution for real-world problems. Therefore, various techniques for aggregating time, products, or resources have been proposed (cf. [18, 19] for an overview).

A mixed-integer linear programming model for mid-term planning of batch production of chemical commodities in a multi-site supply chain is presented by [36]. In this model, early periods are shorter than late periods, which enables an accurate representation of production capacity. A model for a similar problem, but for continuous production mode is formulated by [7].

Perishability of intermediates or final products is vital to the production of fresh food. In [26] the literature about production planning for perishable products is reviewed, and based on an industrial case study of yoghurt production, various mixed-integer linear programming models are developed.

In general, multi-purpose plants are used for batch production. Each product family requires a specific plant configuration, whose setup may take a considerable amount of time. The trade-off between loss of capacity during plant re-configuration and inventory holding cost is addressed by **campaign planning**. General mixed-integer linear programming models for single-site campaign planning are presented by [34]. Inspired by case studies from chemical and pharmaceutical industry, respectively, [18, 19] develop aggregation and hierarchical decomposition techniques for multi-site campaign planning.

All those contributions can be seen as top-down approaches, i.e., the results of the mid-term planning are targets for the subsequent short-term planning. In contrast, [2] propose a bottom-up approach for the coordination of production schedules across multiple plants operating in batch production mode. Here, a **collaborative planning** tool is applied to a large real-world supply chain from chemical industry.

The approaches above assume deterministic input data. Due to changes in input data, mid-term planning should be incrementally updated. Some papers explicitly address **uncertainty**, in particular uncertain demand (cf. e.g. [10]).

4 Short-Term Planning

In this section, we consider short-term planning of production and order acceptance decisions; for procurement and distribution planning confer [11].

Short-term planning of production is performed individually for each plant and is concerned with the detailed allocation of the production equipment and the manpower over time to the production of the primary requirements from mid-term planning. The planning horizon covers at least one period of mid-term planning. The objective is to minimize a (user-defined) combination of makespan, earliness and tardiness with respect to given due dates, and setup cost. Other cost, e.g.

inventory cost, are not considered due to the short planning horizon. In the process industries, typical restrictions are the scarcity of processing units, intermediates, and storage capacities, quarantine and shelf life times for intermediates, general (including cyclic) material flows, flexible proportions of input and output products, and sequence-dependent changeovers and cleanings of processing units or storage facilities (cf. [23]).

In batch production mode, the total requirements for intermediates and final products are divided into batches. To produce a batch, the inputs are loaded into a processing unit, then a transformation process is executed, and finally the output is transferred into storage facilities. The minimum and the maximum filling level of the processing unit used give rise to a lower and an upper bound on the batch size; thus, in general a transformation process has to be executed several times in order to fulfill the primary requirements. Most solution approaches address the problem of short-term planning of batch production as a whole, starting from a mixed-integer formulation (cf. [9, 29] for reviews). In these models, the time horizon is divided into a given number of periods, where the period length is either fixed or variable. In both cases, CPU time requirements for solving real-world problems tend to be prohibitively high. To overcome this difficulty, different heuristics reducing the number of variables have been developed, cf. e.g. [6]. Several successful alternative approaches (e.g. [8, 28, 30]) are based on a decomposition into a planning and a scheduling problem, the latter being solved by problem-specific methods.

In continuous production mode the processing time and the production rate are to be chosen in addition to the proportions of input and output products. Much less work has been reported on short-term planning of continuous production than for batch production; for an overview we refer to [15]. Several authors (e.g. [17, 20]) address the planning of make-and-pack production, which is a typical example of a hybrid production mode: the production system consists of two stages. In the first stage, batches of intermediates are produced. In the second stage, these batches are continuously packed into final products. Often the two stages are separated by some storage facilities.

Order acceptance deals with the on-line decision whether or not accepting new customer orders. The goal is to use efficiently, but not to overload the production facilities. A regression-based solution approach tailored to batch production is presented in [22].

References

1. Applequist G, Samikoglu O, Pekny K, Reklaitis GV (1997) Issues in the use, design and evolution of process scheduling and planning systems. *ISA Transactions* 36:81–121
2. Berning G, Brandenburg M, Gürsoy K, Kussi JS, Mehta V, Tölle FJ (2004) Integrating collaborative planning and supply chain optimization for the chemical process industry (I)—methodology. *Computers and Chemical Engineering* 28:913–927
3. Blackstone JH, Cox JF (2002) APICS dictionary. APICS
4. Blau G, Mehta B, Bose S, Pekny J, Sinclair G, Keunkler K, Bunch P (2000) Risk management in the development of new products in highly regulated industries. *Computers and Chemical Engineering* 24:659–664

5. Blau GE, Pekny JF, Varma VA, Bunch PR (2004) Managing a portfolio of interdependent new product candidates in the pharmaceutical industry. *Journal of Product Innovation Management* 21:227–245
6. Blömer F, Günther HO (2000) LP-based heuristics for scheduling chemical batch processes. *International Journal of Production Research* 38:1029–1051
7. Bok JK, Grossmann IE, Park S (2000) Supply chain optimization in continuous flexible process networks. *Industrial Engineering and Chemistry Research* 39:1279–1290
8. Brucker P, Hurink J (2000) Solving a chemical batch scheduling problem by local search. *Annals of Operations Research* 96:17–38
9. Burkhard RE, Hatzl J (2005) Review, extensions and computational comparison of MILP formulations for scheduling batch processes. *Computers and Chemical Engineering* 29:1752–1769
10. Chen CL, Lee WC (2004) Multi-objective optimization of multi-echelon supply chain networks with uncertain product demands and prices. *Computers and Chemical Engineering* 28:1131–1144
11. Chopra S, Meindl P (2007) *Supply Chain Management*. Prentice Hall
12. Dietz A, Azzaro-Pantel C, Pibouleau L, Domenech S (2006) Multiobjective optimization for multiproduct batch plant design under economic and environmental considerations. *Computers and Chemical Engineering* 30:599–613
13. Flapper SDP, Fransoo JC, Broekmeulen RACM, Inderfurth K (2002) Planning and control of rework in the process industries: a review. *Production Planning & Control* 13:26–34
14. Fleischmann B, Meyr H, Wagner M (2005) Advanced planning. In: Stadtler H, Kilger C (eds) *Supply Chain Management and Advanced Planning*. Springer
15. Floudas CF, Lin X (2004) Continuous-time versus discrete-time approaches for scheduling of chemical processes: a review. *Computers and Chemical Engineering* 28:2109–2129
16. Fransoo JC, Rutten W (1993) A typology of production control situations in process industries. *International Journal of Operations & Production Management* 14:47–57
17. Fündeling CU, Trautmann N (2007) Belegungsplanung einer Make&Pack-Anlage mit kontinuierlicher Prozessführung: eine Fallstudie. In: Günther HO, Mattfeld DC, Suhl L (eds) *Supply Network and Logistics Management*. Physica
18. Grunow M, Günther HO, Lehmann M (2002) Campaign planning for multi-stage batch processes in the chemical industry. *OR Spectrum* 24:281–314
19. Grunow M, Günther HO, Yang G (2003) Plant co-ordination in pharmaceuticals supply networks. *OR Spectrum* 25:109–141
20. Günther HO, Grunow M, Neuhaus U (2006) Realizing block planning concepts in make-and-pack production using MILP modelling and SAP APO. *International Journal of Production Research* 44:3711–3726
21. Guirardello R, Swaney RE (2005) Optimization of process plant layout with pipe routing. *Computers and Chemical Engineering* 30:99–114
22. Ivănescu VC, Fransoo JC, Bertrand JWM (2006) A hybrid policy for order acceptance in batch process industries. *OR Spectrum* 28:199–222
23. Kallrath J (2002a) Planning and scheduling in the process industry. *OR Spectrum* 24:219–250
24. Kallrath J (2002b) Combined strategic and operational planning – an MILP success story in chemical industry. *OR Spectrum* 24:315–341

25. Kallrath J (2005) Solving planning and design problems in the process industry using mixed integer and global optimization. *Annals of Operations Research* 140:339–373
26. Lütke Entrup M, Günther HO, van Beek P, Grunow M, Seiler T (2005) Mixed-integer linear programming approaches to shelf-life-integrated planning and scheduling in yoghurt production. *International Journal of Production Research* 43:5071–5100
27. Maravelias CT, Grossmann IE (2004a) Optimal resource investment and scheduling of tests for new product development. *Computers and Chemical Engineering* 28:1021–1038
28. Maravelias CT, Grossmann IE (2004b) A hybrid MILP/CP decomposition approach for the continuous time scheduling of multipurpose batch plants. *Computers and Chemical Engineering* 28:1921–1949
29. Méndez CA, Cerdá J, Grossmann IE, Harjunkoski I, Fahl M (2006) State-of-the-art review of optimization methods for short-term scheduling of batch processes. *Computers and Chemical Engineering* 30:913–946
30. Neumann K, Schwindt C, Trautmann N (2002) Advanced production scheduling for batch plants in process industries. *OR Spectrum* 24:251–279
31. Schmidt CW, Grossmann IE, Blau GE (1998) Optimization of industrial scale scheduling problems in new product development. *Computers and Chemical Engineering* 22:S1027–S1030
32. Shah N (2004) Pharmaceutical supply chains: key issues and strategies for optimisation. *Computers and Chemical Engineering* 28:929–941
33. Shah N (2005) Process industry supply chains: advances and challenges. *Computers and Chemical Engineering* 29:1225–1235
34. Suerie C (2005) Campaign planning in time-indexed model formulations. *International Journal of Production Research* 43:49–66
35. Taylor SG, Seward SM, Bolander SF (1981) Why the process industries are different. *Production and Inventory Management Journal* 22:9–24
36. Timpe CH, Kallrath J (2000) Optimal planning in large multi-site production networks. *European Journal of Operational Research* 126:422–435
37. Tsiakis P, Shah N, Pantelides CC (2001) Design of multi-echelon supply chain networks under demand uncertainty. *Industrial and Engineering Chemistry Research* 40:3585–3604
38. Wouda FHE, van Beek P, van der Vorst JGAJ, Tacke H (2002) An application of mixed-integer linear programming models on the redesign of the supply network of Nutricia Dairy & Drinks Group in Hungary. *OR Spectrum* 24:449–465

Production Planning in Dynamic and Seasonal Markets

Jutta Geldermann, Jens Ludwig, Martin Treitz, and Otto Rentz

French-German Institute for Environmental Research, University of Karlsruhe,
Germany
`jutta.geldermann@wiwi.uni-karlsruhe.de`

Summary. In chemical engineering, pinch analysis holds a long tradition as a method for determining optimal target values for heat or mass exchanger networks by calculating an optimal alignment of available flows. A graphical representation of the time-material production relationship derived from the original pinch analysis can be used for aggregate production planning. This can deliver insights in the production planning problem as several production strategies can be compared. The approach is applied to a case study on bicycle coating.

1 Introduction

The problem of planning the future production capacity of a company can be addressed in numerous ways and has been extensively discussed in literature. A good overview of standard methods is given in [4]. For more recent approaches, the reviews of [5] and [9] are recommended, with the latter focusing on high-tech industries. When companies face a seasonal demand, the problem of capacity adaptation can become even more challenging, as there may be a constant need for capacity adaptations. Forecasting such seasonal demands is possible using various statistical methods, mostly aiming at identifying a seasonal component in historical demand data (see for example [8]). Given a prediction of the upcoming seasonal demand, a company still has to choose its production strategy, i.e., when to operate at which production rate. [6, 7] propose a graphical method that represents demand and supply data as composite curves and derives inspiration of pinch analysis. This approach is a classical method from the chemical process industry that aims at optimizing a system's performance by analyzing the process streams, i.e. mass and energy flows, and possible interconnections. The same methodology can be applied to product streams and the time-material production relationship. The application of the analysis is shown for a Chinese bicycle company facing seasonal changes throughout the year and different production planning strategies are compared based on cost criteria. In this application, the dependence of the planning outcome on the starting season and the occurrence of stock-out using the pinch planning method were identified as shortcomings of the method for which solutions are proposed.

2 Production Planning Applying the Pinch Analysis Approach

The Classical Pinch Analysis for Heat Integration

The basic idea of the thermal pinch analysis is a systematic approach to the minimisation of lost energy in order to come as close as possible to a reversible system (see [3, 2]). In its first step the pinch analysis yields the best possible heat recovery at the thermodynamic optimum. Thus, the pinch analysis requires the combination of hot and cold process streams to composite curves and the description of the respective temperature-enthalpy relationships. However, there exists a trade-off between the savings in operating costs for the hot utility and the investment in the heat exchanger. The result of the pinch analysis is the energy savings potential for the considered set of processes.

Translation of the Thermal Pinch Analysis to Production Planning

An analysis of intra- and inter-company production networks on the basis of product streams is also possible in analogy to the classical pinch analysis. The *time-material production* relationship can be used for a pinch analysis approach for aggregate production planning [6, 7]. Based on material balances, a time versus material quantity plot can be derived in translation of the original thermal pinch analysis. The quality parameter in the production planning pinch is the time of production (in analogy to the temperature level T). The quantity parameter is the demand of units to a certain time (in analogy to the enthalpy ΔH , describing the sum of internal energy of a thermodynamic system. One demand composite curve and one production composite curve can be constructed on a time basis (cf. Figure 1).

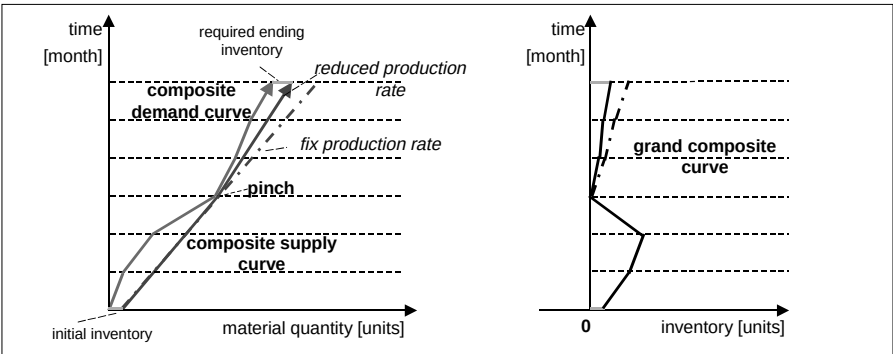


Fig. 1. Production Planning [6]

In this context aggregated production planning is defined as the identification of an overall level of production for an individual company. The focus of the analysis is the evaluation of seasonal changes on the demand side and its consequences for setting the level of production during the whole period considered. The central issue

in this case is how to choose and adapt the production rate during the period in order to avoid stock-outs and minimize inventory and capacity changes.

3 Discussion of Production Strategies

Different options for the level of production can be discussed based on costs, such as investment dependent costs, labour costs, material costs, inventory costs, stock-out penalty costs and costs for capacity adjustments (hiring and layoff costs). The strategies illustrate different ways based on flexibility and costs to comply with the demand composite curve and to supply the required aggregated demand (cf. Table 1). *Variable Production Level with one Pinch Point* is the strategy derived from pinch analysis by [6, 7]. It consists of choosing the minimal production rate from the starting point that does not create stock-outs. The point where this straight line is tangent to the demand composite curve is called pinch point and the production rate is adapted here in break with its ending in the ending point required by the demand composite curve.

However, this strategy generated stock-outs for some demand patterns, especially if no ending inventory is planned. In case such stock-outs occur, we propose two corrective strategies: Firstly, the higher production rate before the pinch point could be maintained for as many periods, as are needed in order to build up sufficient inventory levels. This solution increases inventory costs, but passes on additional adjustments of the production rate.

A second solution is strategy 3 called *Variable Production Level with multiple Pinch Points*, which foresees an adaptation of the production rate in several pinch points, determined as the point of contact of the minimal production rate below the demand composite curve, starting from the last pinch point. This strategy may entail several adaptations of the production rate, however only in cases where the one pinch point strategy leads to stock-outs. As this strategy is viable for most given demand patterns, it was superior to the other strategies in most cases. This strategy could be further enhanced, for example by considering the trade-off between stock-outs, inventory and capacity adaptations for every pinch point. The *Average Production* strategy and the *Max-Zero* strategy are used as benchmarks for the evaluation of the trade-off of the penalty for stock-outs on the one hand and the inventory costs on the other.

Table 1. Available Strategies in the Production Pinch Analysis

No.	Strategy	Pinch Points	Capacity Adjustments	Possibility Stock-Out
1	Fixed Production Level with one Pinch Point	1	0	no
2	Variable Production Level with one Pinch Point	1	1	no
3	Variable Production Level with multiple Pinch Points	var.	var.	no
4	Average Production	0	0	yes
5	Max-Zero Strategy	0	1	no

Challenge for the analysis is the determination of the starting time interval for the analysis. In contrast to the thermal pinch analysis in which all heating and cooling requirements are sorted according to their quality parameter temperature and resulting a theoretical minimal utility target, the sorting of the demand in the production pinch analysis is infeasible. Consequently, the analysis can result in sub-optimal solutions and the results significantly alter depending on the selected starting interval. Here, we propose the beginning of the peak season, which means the highest growth rates, as the starting point of the evaluation. However, more research is needed here in order to provide recommendations for arbitrary demand lines.

4 Case Study

In the following the bicycle production of a reference company in China is taken into account (see [1]) and a distribution of the demand for bicycles throughout the year is assumed according to Table 2. The five basic strategies for production planning are discussed (cf. Table 1). Since the seasonal increase of demand starts in October it is the starting month of the evaluation.

Table 2. Monthly Production Rates and Estimated Costs for all Strategies

Demand		Production Rates [Bikes/Month]				
		Fixed Production with one Pinch-Point	Variable Production with one Pinch-Point	Variable Production, multiple Pinch-Points	Average Production	Max-Zero Strategy
October	90.000	108.000	108.000	108.000	83.333	90.000
November	90.000					90.000
December	110.000					110.000
January	130.000					130.000
February	120.000					120.000
March	90.000			90.000		90.000
April	80.000			80.000		80.000
May	60.000		48.800	60.000	0	60.000
June	50.000			50.000		50.000
July	50.000			50.000		50.000
August	60.000			60.000		60.000
September	70.000			70.000		70.000
Annual Cost [Thousands €]		30.918	22.539	21.702	23.465*	37.892
Ending Inventory [Bikes]		296.000	0	0	0	0

Table 2 shows the demand for each regarded month and the corresponding production rate according to the five strategies. The costs of each production strategy are calculated using estimated values for material, labour and inventory. Relevant costs for choosing the production strategy are hiring and layoff costs on one hand and inventory and stock-out costs on the other hand. In this example, the strategies inspired by pinch analysis achieve minimal costs, meaning they offer the best compromise between the objectives of low inventory/stock-out and a small number of costly capacity adaptations.

Strategy 3 (lowest costs) and 5 (highest costs) are shown in Figure 2 taking into account a safety stock ΔI_{min} of 5 000 units. In Strategy 3 multiple capacity adjustments are allowed compared to Strategy 2 with only one capacity adjustment. As Figure 2 illustrates the production curve closely follows the demand curve by identifying four pinch points. However, the determination of the starting time interval of the evaluation highly influences the results of the analysis. For the analysis the beginning of the peak season is proposed as the starting interval of the analysis. More comprehensive research concerning the time-variance of the results is necessary.

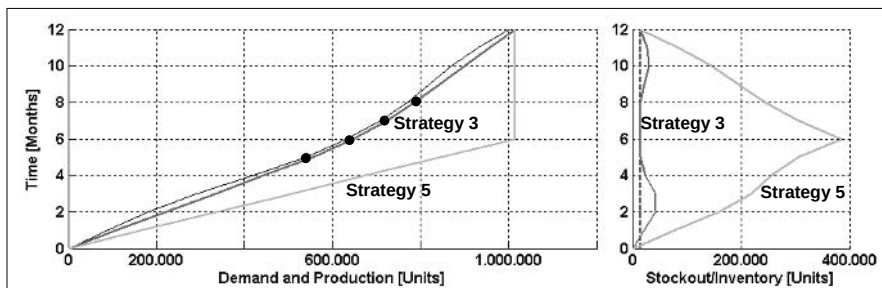


Fig. 2. Composite and Grand Composite Curves of the Production Pinch with Strategies 3 and 5

5 Conclusions

The application of the pinch analysis methodology to production planning provides a simple but effective tool for analysing production strategies of a company in a dynamic market. Based on the cost parameters the different production strategies can be evaluated, those based on pinch analysis provide a good compromise between inventory costs and capacity adaptation costs.

The basic idea by [6, 7] has been improved by an opening step comprising the choice of a beginning period with the highest demands. Moreover, different strategies have been introduced. The idea to select a strategy with multiple pinch points was proposed in order to overcome stock-outs possible in the one pinch point strategy. Especially this strategy seems to be promising. Further investigations are necessary as regards the shape of the demand curves suited for the application of this method.

Acknowledgement. This research is funded by a grant from the *VolkswagenStiftung*. We would like to thank them for the excellent support of our research.

References

1. J. Geldermann, H. Schollenberger, M. Treitz, O. Rentz (2006) Adapting the european approach of best available techniques: Case studies from chile and china. *Journal of Cleaner Production* (accepted)
2. J. Geldermann, M. Treitz, O. Rentz (2006) Integrated technique assessment based on the pinch analysis approach for the design of production networks. *European Journal of Operational Research* 171(3): 1020–1032
3. B. Linnhoff, D. Manson, I. Wardle (1979) Understanding heat exchanger networks. *Computers and Chemical Engineering* 3: 295–302
4. H. Luss (1982) Operations research and capacity expansion problems: A survey. *Operations Research*, 30(5): 907–947
5. J. Van Mieghem (2003) Capacity management, investment, and hedging: Review and recent developments. *Manufacturing & Service Operations Management*, 5(4): 269–302, 2003
6. A. Singhvi, U. V. Shenoy (2002) Aggregate Planning in Supply Chains by Pinch Analysis. *Transaction of the Institution of Chemical Engineers*, 80(A): 2002
7. A. Singhvi, K. P. Madhavan, and U. V. Shenoy (2004) Pinch analysis for aggregate production planning in supply chains. *Computers and Chemical Engineering*, 28: 993–999
8. P. Winters (1960) Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3): 324–342
9. S. Wu, M. Erkoç, S. Karabuk (2005) Managing capacity in the high-tech industry: A review of literature. *Engineering Economist*, 50(2): 125–158

A Branch and Bound Algorithm Based on DC Programming and DCA for Strategic Capacity Planning in Supply Chain Design for a New Market Opportunity

Nguyen Canh Nam¹, Le Thi Hoai An², and Pham Dinh Tao¹

¹ Laboratory of Modelling, Optimization & Operations Research (LMI). National Institute for Applied Sciences - Rouen, France
nguyencn@insa-rouen.fr
pham@insa-rouen.fr

² Laboratory of Theoretical and Applied Computer Science (LITA). Informatics Department UFR MIM, Metz University, France
lethi@sciences.univ-metz.fr

1 Introduction

Supply chain network are considered as solution for effectively meeting customer requirements such as low cost, high product variety, quality and shorter lead times. The success of a supply chain lies in good strategic and tactical planning and monitoring at the operational level. Strategic planning is long term planning and usually consists in selecting providers and distributors, location and capacity planning of manufacturing/servicing units, among others.

In this paper we only consider the problem of strategic capacity planning proposed in [5, 6]. We consider a three-echelon system : providers, producers and distributors. We assume that all partners under consideration have already met the pre-requisite requirements. All distributors have a definite demand in each period on a given horizon. Providers provide the raw material/semi-finished products to selected producers and these ones fulfill the demand of each distributor. Each provider and producer has its production cost and transportation cost to the next stage. These costs are invariant in time. The transportation costs vary from one pair (provider-producer and producer-distributor) to another. Each provider and each producer has limited production and transportation capacities. The transportation and production capacities can be extended by investing in resource. We assume that investment is only possible at the beginning of the first period. The transportation costs and the production costs are linear functions of quantities. Investment cost is also a linear function of quantity, but with additional fixed cost which does not depend on quantities but only on the entities it is related to. The problem is to select the most economic combination of providers and producers such that they satisfy the demand imposed by all distributors.

This problem is modeled as a mixed-integer program as follows. Let $i \in \{1, 2, \dots, P\}$ be the providers, $j \in \{1, 2, \dots, M\}$ be the producers and $k \in \{1, 2, \dots, D\}$ be the distributors. The demand is deterministic and is known for each distributor on time horizon T . We denote

- p_i : Raw material cost per unit at provider i .
- m_j : Production cost per unit at producer j .
- u_{ij} : Transportation cost per unit from provider i to producer j .
- v_{jk} : Transportation cost per unit from producer j to distributor k .
- R_{ij}, r_{ij} : Available, added transportation capacities from i to j .
- S_{jk}, s_{jk} : Available, added transportation capacities from j to k .
- G_i, g_i : Available, added production capacities for provider i .
- H_j, h_j : Available, added production capacities for producer j .
- d_k^t : Demand of distributor k in period t .

We consider four investments in this model : investment to enhance the capacity of providers, investment to enhance the capacity of producers and investment to enhance the transportation capacity between provider-producer and producer-distributor. So we have the following four investment cost :

- Investment cost for new transportation capacity r_{ij} from provider i to producer j : Cost = $\mathbf{A}_{ij} + \mathbf{a}_{ij}\mathbf{r}_{ij}$ if $r_{ij} > 0$ and equal to $\mathbf{0}$ if $r_{ij} = 0$
- Investment cost for new transportation capacity s_{jk} from producer j to distributor k : Cost = $\mathbf{B}_{jk} + \mathbf{b}_{jk}\mathbf{s}_{jk}$ if $s_{jk} > 0$ and equal to $\mathbf{0}$ if $s_{jk} = 0$
- Investment cost for provider i to enhance its capacity by g_i : Cost = $\mathbf{E}_i + \mathbf{e}_i\mathbf{g}_i$ if $g_i > 0$ and equal to $\mathbf{0}$ if $g_i = 0$
- Investment cost for producer j to enhance its capacity by h_j : Cost = $\mathbf{F}_j + \mathbf{f}_j\mathbf{h}_j$ if $h_j > 0$ and equal to $\mathbf{0}$ if $h_j = 0$

Finally, we have the model with the objective function

$$\begin{aligned}
 & \sum_{t=1}^T \sum_{i=1}^P \sum_{j=1}^M C_t(p_i + u_{ij})x_{ij}^t + \sum_{t=1}^T \sum_{j=1}^M \sum_{k=1}^D C_t(m_j + v_{jk})y_{jk}^t + \\
 & \sum_{i=1}^P \sum_{j=1}^M (C_1 - D_1)(a_{ij}r_{ij} + A_{ij}U_{ij}) + \sum_{j=1}^M \sum_{k=1}^D (C_1 - D_1)(b_{jk}s_{jk} + B_{jk}V_{jk}) + \\
 & \sum_{i=1}^P (C_1 - D_1)(e_i g_i + E_i W_i) + \sum_{j=1}^M (C_1 - D_1)(f_j h_j + F_j X_j) \quad (1)
 \end{aligned}$$

where $C_t = (1 + \alpha)^{T-t+1}$, $D_t = (1 - \beta)^{T-t+1} \quad \forall t \in \{1, 2, \dots, T\}$ with α is the discount rate and β is the depreciation factor.

and the constraints :

$$\sum_{j=1}^M x_{ij}^t \leq G_i + g_i, \quad x_{ij}^t \leq R_{ij} + r_{ij} \quad t \in \overline{1, T}, i \in \overline{1, P}, j \in \overline{1, M} \quad (2)$$

$$\sum_{k=1}^D y_{jk}^t \leq H_j + h_j, \quad y_{jk}^t \leq S_{jk} + s_{jk} \quad t \in \overline{1, T}, j \in \overline{1, M}, k \in \overline{1, D} \quad (3)$$

$$\sum_{i=1}^P x_{ij}^t = \sum_{k=1}^D y_{jk}^t, \quad \sum_{j=1}^M y_{jk}^t = d_k^t \quad t \in \overline{1, T}, j \in \overline{1, M}, k \in \overline{1, D} \quad (4)$$

$$r_{ij} \leq Z_{ij}^r U_{ij} \quad i \in \overline{1, P}, j \in \overline{1, M}; \quad s_{jk} \leq Z_{jk}^s V_{jk} \quad j \in \overline{1, M}, k \in \overline{1, D} \quad (5)$$

$$g_i \leq Z_i^g W_i \quad i \in \overline{1, P}; \quad h_j \leq Z_j^h X_j \quad j \in \overline{1, M} \quad (6)$$

$$U_{ij} \in \{0, 1\}, U_{jk} \in \{0, 1\}, W_i \in \{0, 1\}, X_j \in \{0, 1\} \quad (7)$$

where $Z_{jk}^s, Z_{ij}^r, Z_i^g, Z_j^h$ are big numbers greater than the total demand.

The constraints (2) and (3) guarantee that the shipment and delivery must not exceed the available transportation and production capacity respectively. Constraints (4) are balance constraints. The binary variables U_{ij}, V_{ij}, W_i and X_j are associated with fixed costs in objective function and take positive values only if capacities enhancements are made.

We see that this problem is in the form of a large-scale mixed integer linear program. Most researches in literature for this problem are based on heuristic [5, 6]. In this work, we present a method, for solving this problem, in which we combine the local algorithm DCA in DC programming with the adapted Branch and Bound for globally solving the class of mixed 0-1 linear programs [2]. To this purpose, we first apply the theory of exact penalization in DC programming developed by Le Thi Hoai An et al [1] to reformulate Problem (1) as that of minimizing a DC function over a polyhedral convex set. The resulting problem is then handled by DCA in DC programming approach introduced by Pham Dinh Tao in 1985 and extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao (see [7, 8, 4, 3] and references therein).

The article is organized as follows. An introduction is followed by a brief description of DCA in the section 2, where the DC reformulation and its solution algorithm DCA are established. The last section reports computational results on a series of test problems.

2 DC Programming Approach

2.1 DC Reformulation

In this section, by using an exact penalty result we will formulate (1) in the form of a concave minimization program. Let $x \in \mathbb{R}^n$ be the continuous variables and $y \in \mathbb{R}^m$ be the binary variables in (1). Clearly, the set S of feasible points (x, y) determined by $\{(2), \dots, (6)\}$ is a nonempty, bounded polyhedral convex set in $\mathbb{R}^n \times \mathbb{R}^m$. Problem (1) can be expressed in the form:

$$\alpha = \min : \{c^T x + d^T y : (x, y) \in S, y \in \{0, 1\}^m\}. \quad (8)$$

Let us consider the function p defined by $p(x, y) = \sum_{i=1}^m \min\{y_i, 1 - y_i\}$.

Set $K := \{(x, y) \in S : y \in [0, 1]^m\}$. It is clear that p is concave and finite on K , $p(x, y) \geq 0$ for all $(x, y) \in K$, and

$$\{(x, y) \in S : y \in \{0, 1\}^m\} = \{(x, y) \in K : p(x, y) \leq 0\}.$$

Hence Problem (8) can be rewritten as

$$\alpha = \min : \{c^T x + d^T y : (x, y) \in K, p(x, y) \leq 0\} \quad (9)$$

From Theorem 1 below we get, for a sufficiently large number t ($t > t_0$), the equivalent concave minimization problem to (8) :

$$\min : \{c^T x + d^T y + tp(x, y) : (x, y) \in K\}. \quad (10)$$

Theorem 1. (Theorem 1, [1]) *Let K be a nonempty bounded polyhedral convex set, f be a finite concave function on K and p be a finite nonnegative concave function on K . Then there exists $t_0 \geq 0$ such that for $t > t_0$ the following problems have the same optimal value and the same solution set:*

$$\begin{aligned} (P_t) \quad \alpha(t) &= \inf \{f(x) + tp(x) : x \in K\} \\ (P) \quad \alpha &= \inf \{f(x) : x \in K, p(x) \leq 0\}. \end{aligned}$$

More precisely if the vertex set of K , denoted by $V(K)$, is contained in $\{x \in K : p(x) \leq 0\}$, then $t_0 = 0$, otherwise $t_0 = \min \left\{ \frac{f(x) - \alpha(0)}{S} : x \in K, p(x) \leq 0 \right\}$, where $S := \min \{p(x) : x \in V(K), p(x) > 0\} > 0$. $\square$

Clearly that (10) is a DC program with

$$g(x, y) = \chi_K(x, y) \quad \text{and} \quad h(x, y) = -c^T x - d^T y - t \sum_{i=1}^m \min\{y_i, 1 - y_i\} \quad (11)$$

where $\chi_K(x, y) = 0$ if $(x, y) \in K$, 0 otherwise (the indicator function of K).

2.2 DCA for Solving Problem (10)

In this section we investigate a DC programming approach for solving (??). A DC program is that of the form

$$\alpha := \inf \{f(x) := g(x) - h(x) : x \in \mathbb{R}^n\}, \quad (12)$$

with g, h being lower semicontinuous proper convex functions on $\mathbb{R}^n$, and its

$$\text{dual is defined as} \quad \alpha := \inf \{h^*(y) - g^*(y) : y \in \mathbb{R}^n\}, \quad (13)$$

where $g^*(y) := \sup \{\langle x, y \rangle - g(x) \mid x \in \mathbb{R}^n\}$, the conjugate function of g .

Based on local optimality conditions and duality in DC programming, the DCA consists in the construction of two sequences $\{x^k\}$ and $\{y^k\}$, candidates to be optimal solutions of primal and dual programs respectively, in such a way that $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing and their limits points satisfy the local optimality conditions. It turns out that DCA's scheme takes the simple form $\mathbf{y}^k \in \partial h(\mathbf{x}^k)$; $\mathbf{x}^{k+1} \in \partial g^*(\mathbf{y}^k)$.

It should be reminded that if either g or h is polyhedral convex, then (12) is called a polyhedral DC program for which DCA has a finite convergence ([7, 8, 4]). That is the case for DCA applied to (10).

DCA for solving (10).

By the very definition of h , a subgradient $(u, v) \in \partial h(x, y)$ can be chosen :

$$(u, v) \in \partial h(x, y) \Leftarrow u = -c, \quad v = -d + \begin{cases} t & y_i \geq 0.5 \\ -t & \text{otherwise.} \end{cases} \quad (14)$$

Algorithm 1 (DCA applied to (10))

Let $\epsilon > 0$ be small enough and (x^0, y^0) . Set $k \leftarrow 0$; $er \leftarrow 1$.
while $er > \epsilon$ **do**
 Compute $(u^k, v^k) \in \partial h(x^k, y^k)$ via (14).
 Solve the linear program: $\min\{-(\langle u^k, v^k \rangle, (x, y)) : (x, y) \in K\}$
 to obtain (x^{k+1}, y^{k+1})
 $er \leftarrow ||(x^{k+1}, y^{k+1}) - (x^k, y^k)||$
 $k \leftarrow k + 1$
endwhile

The convergence of Algorithm 1 can be summarized in the next theorem whose proof is essentially based on the general convergence of DCA applied to polyhedral DC program ([7, 4, 2]).

Theorem 2. (Convergence properties of Algorithm 1)

- (i) Algorithm 1 generates a sequence $\{(x^k, y^k)\}$ contained in $V(K)$ such that the sequence $\{g(x^k, y^k) - h(x^k, y^k)\}$ is decreasing.
- (ii) If at iteration r we have $y^r \in \{0, 1\}^m$, then $y^k \in \{0, 1\}^m$ for all $k \geq r$.
- (iii) The sequence $\{(x^k, y^k)\}$ converges to $(x^*, y^*) \in V(K)$ after a finite number of iterations. The point (x^*, y^*) is a critical point of Problem (10) (i.e. $\partial g(x^*, y^*) \cap \partial h(x^*, y^*) \neq \emptyset$). Moreover if $y_i^* \neq \frac{1}{2}$ for $i = 1, \dots, m$, then (x^*, y^*) is a local solution to Problem (10).

3 Computational Experiments

To prove globality of solutions computed by DCA and improve the adapted Branch and Bound (BB) [2], we consider also the combined DCA-Branch and Bound (DCA-BB). These algorithms were implemented on the Sony computer in C++. To solve related linear programs, we used software CPLEX version 7.5. For testing the performance of these algorithms, we randomly generated 10 examples, according to the scheme in [5, 6], with number of entries varying from 4 to 10 for each echelon and number of periods varying from 2 to 10 in the planning horizon (sizes already considered as large in [6, 5]). For all test problems, we always obtained ϵ -optimal solutions, with $\epsilon \leq 0.05$. Moreover, our method DCA provides best upper bounds after a very few restartings from initial points given by the combined algorithm. In Table 1 we compare the performance of the two algorithms BB and DCA-BB.

Table 1. Numerical results

Pr	B&B				DCA - B&B			
	It	UB	LB	t(s)	It	UB	LB	t(s)
P1	24	692644.25	691105.58	0.83	4	692395.86	676336.78	0.44
P2	71	2884757.52	2884757.52	3.80	5	2928008.48	2831637.44	0.83
P3	98	773552.73	773464.64	4.13	9	775421.07	758579.62	1.27
P4	35	55498904.62	54356906.11	2.64	4	54570240.95	52787572.92	0.97
P5	68	7781457.67	7774747.61	3.53	6	7801039.33	7605506.39	1.00
P6	29	5356292.69	5356292.69	1.63	5	5356292.69	5217137.62	0.83
P7	33	34007298.98	34004351.21	2.82	4	34031846.80	33470976.11	0.88
P8	6	32800300.89	32555249.45	0.45	1	32557626.05	32140930.39	0.22
P9	30	5464826.08	5401633.50	2.08	2	5503654.90	5258447.02	0.41
P10	50	6046483.77	6007710.54	4.40	4	6008074.95	5904890.75	1.09

Conclusion. We have presented a DC programming approach and DCA for solving the Supply Chain Design Problem in Strategic Capacity Planning. The method is based on DCA and Branch and Bound technique. The results show that the DCA is very efficient and original as it can give integer solutions while working in a continuous domain.

References

1. Le Thi Hoai An, Pham Dinh Tao and Le Dung Muu (1999) Exact Penalty in DC Programming. *Vietnam Journal of Mathematics* 27:2, pp 1216–1231.
2. Le Thi Hoai An, Pham Dinh Tao (2001) A continuous approach for globally solving linearly constrained 0-1 programming problems, *Journal of Global Optimization* 50, pp. 93-120.
3. Le Thi Hoai An, Pham Dinh Tao (2003) Large Scale Molecular Optimization From Distance Matrices by a DC Optimization Approach. *SIAM Journal on Optimization*, Vol. 14, No 1 (2003), pp. 77-116.
4. Le Thi Hoai An, Pham Dinh Tao (2005) The DC (Difference of Convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problem. *Annals of Operations Research* 133 : 23–46.
5. Satyaveer S.Chauhan, Rakesh Nagi and Jean-Marie Proth (2002) Strategic Capacity Planning in Supply Chain Design for a New Market Opportunity. INRIA Research Report, RR-4658.
6. Satyaveer S.Chauhan (2003) Chaînes d'approvisionnement : approches stratégique et tactique. Thèse de Doctorat, Université de Metz.
7. Pham Dinh Tao, Le Thi Hoai An (1997) Convex analysis approach to DC programming : Theory, Algorithms and Applications. *Acta Mathematica Vietnamica*, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday 22:1, pp 289–355.
8. Pham Dinh Tao and Le Thi Hoai An (1998) DC optimization algorithms for solving the trust region subproblem. *SIAM J.Optimization* 8:476–505.

Scheduling and Project Management

Branching Based on Home-Away-Pattern Sets

Dirk Briskorn and Andreas Drexl

Lehrstuhl für Produktion & Logistik, Christian-Albrechts-Universität zu Kiel,
Germany

`briskorn@bwl.uni-kiel.de`

`andreas.drexl@bwl.uni-kiel.de`

1 Introduction

Scheduling a sports league requires to solve a hard combinatorial optimization problem. We consider a league of a set T of n teams supposed to play in a single round robin tournament (SRRT). Accordingly, each team $i \in T$ has to play against each other team $j \in T, j \neq i$, exactly one match. The tournament is partitioned into matchdays (MD) being periods where matches can be carried out. Each team $i \in T$ shall play exactly once per MD. Hence, we have a compact structure resulting in an ordered set P of $n - 1$ MDs.

Since each match has to be carried out at one of both opponents' venues breaks come into play. A break for team i occurs at MD p if i has two consecutive matches at home or away, respectively, at MDs $p - 1$ and p . We distinguish between home-breaks and away-breaks depending on the breaks' venue. A popular goal concerning breaks is to minimize the number of their occurrence. Apparently, the number of breaks can not be less than $n - 2$; see [3].

There have been several efforts in order to find good or even optimal SRRTs, see [1] and [4] for example. However, all available algorithms search only a rather small part of the solution space which is a lack we aim to overcome. In section 2 the problem is described in detail and a corresponding mathematical model is given. Section 3 focuses on the branching scheme to tackle the problem. Some computational results are outlined in section 4 and, finally, section 5 gives some conclusions and an outlook to future research.

2 Model

Assume that we have cost $c_{i,j,p}$ if team i plays at home against team j at MD p . The goal of the SRRT introduced in [2] is to find a SRRT having the minimum sum of arranged matches' cost. In this paper, we construct cost-minimal SRRTs while assuring the minimum number of breaks (MBSRRT). The corresponding integer program can be given as follows:

We employ binary match variables $x_{i,j,p}$ being equal to 1 if and only if team i plays at home against team j at MD p . Then, the objective function (1) represents

$$\min \sum_{i \in T} \sum_{j \in T \setminus \{i\}} \sum_{p \in P} c_{i,j,p} x_{i,j,p} \quad (1)$$

$$\text{s.t.} \sum_{p \in P} (x_{i,j,p} + x_{j,i,p}) = 1 \quad \forall i, j \in T, i < j \quad (2)$$

$$\sum_{j \in T \setminus \{i\}} (x_{i,j,p} + x_{j,i,p}) = 1 \quad \forall i \in T, p \in P \quad (3)$$

$$\sum_{j \in T \setminus \{i\}} (x_{i,j,p-1} + x_{i,j,p}) - br_{i,p} \leq 1 \quad \forall i \in T, p \in P^{\geq 2} \quad (4)$$

$$\sum_{j \in T \setminus \{i\}} (x_{j,i,p-1} + x_{j,i,p}) - br_{i,p} \leq 1 \quad \forall i \in T, p \in P^{\geq 2} \quad (5)$$

$$\sum_{i \in T} \sum_{p \in P^{\geq 2}} br_{i,p} \leq n - 2 \quad (6)$$

$$x_{i,j,p} \in \{0,1\} \quad \forall i, j \in T, i \neq j, p \in P \quad (7)$$

$$br_{i,p} \in \{0,1\} \quad \forall i \in T, p \in P^{\geq 2} \quad (8)$$

the goal of cost minimization. Restrictions (2) and (3) form a SRRT by letting each pair of teams meet exactly once and letting each team play exactly once per MD. Restrictions (4) and (5) set the binary break variable $br_{i,p}$ to 1 if team i plays twice at home or twice away at MDs $p-1$ and p . Inequality (6) assures that no more than $n-2$ breaks are arranged.

3 Branching Scheme

Branching must take care of the special structure of the problem at hand, because only few branches are able to assure that the corresponding subtree will have at least one single feasible solution. However, detecting infeasibility in advance is pretty complicated but worth to be done because otherwise a huge computational effort is spent on infeasible subtrees. Consider for example three branches setting variables $x_{i,j,p} = x_{i,j',p+3} = x_{i,j'',p+6} = 1$ for arbitrary pairwise disjoint teams $i, j, j', j'' \in T$ and MD $p \in P^{\leq |P|-6}$. These branches imply that team i has a break at MD $p', p+1 \leq p' \leq p+3$, as well as at MD $p'', p+4 \leq p'' \leq p+6$. However, because each team must not have more than one break in a MBSRRT (see [5]), the corresponding subtree does not permit to construct any feasible solution to the problem at hand.

[5] study the structure of MBSRRTs and proof several characteristics which can be employed in order to avoid searching subtrees having no feasible solution. The basic idea is to branch on break-variables according to rules derived from the four characteristics i to iv provided in [5]. In MBRRTs

- i each team can have no more than one break,
- ii at each MD either two breaks occur or none,
- iii the two teams having no break can be said to have breaks at MD 1, and

- iv following from ii and iii each team i is paired with another team j where i plays at home at MD p if and only if j plays away at p .

From i we conclude that by fixing a break (defined by MD and venue) for a team i the venue for each MD is inherently set for i as well. If we branch on i to have a home-break at MD p then we can fix to zero half the match variables corresponding to i as follows:

$$\begin{aligned} x_{i,j,p'} &= 0 \quad \forall p' \in P, ((p' < p \wedge p - p' \text{ even}) \vee (p' > p \wedge p' - p \text{ odd})) \\ x_{j,i,p'} &= 0 \quad \forall p' \in P, ((p' < p \wedge p - p' \text{ odd}) \vee (p' \geq p \wedge p' - p \text{ even})) \end{aligned}$$

Consequently, fixing variables as a result of setting an away-break for team i at MD p is done the other way round:

$$\begin{aligned} x_{j,i,p'} &= 0 \quad \forall p' \in P, ((p' < p \wedge p - p' \text{ even}) \vee (p' > p \wedge p' - p \text{ odd})) \\ x_{i,j,p'} &= 0 \quad \forall p' \in P, ((p' < p \wedge p - p' \text{ odd}) \vee (p' \geq p \wedge p' - p \text{ even})) \end{aligned}$$

Furthermore, we know from iv that if there is a team i having a home-break (away-break) at MD p there is a team $j \neq i$ having an away-break (home-break) at MD p . Note that there must be two breaks at MD 1 in order to reach the minimum number of breaks. Therefore, we have to choose $\frac{n-2}{2}$ periods out of $P^{\geq 2}$ and assign the resulting $n - 2$ breaks to $n - 2$ pairwise disjoint teams. For the remaining two teams we have to decide which one has a home-break at MD 1 and which one has an away-break at MD 1.

We propose a static branching scheme as follows: for each node k of the branching tree with depth d_k we construct a child by setting one particular possible break for team $d_k + 1$. Obviously, the set of possible breaks and, therefore, the number of the children of k depends on the branches on the path $\vec{V}_k$ from the root node to k . Let $nb_{p, \vec{V}_k}$ be the number of breaks set at MD p on $\vec{V}_k$ and $P_{\vec{V}_k}^{br}$ be the set of MDs where at least one break was set at on $\vec{V}_k$. Then we can state rules I to III to branch at node k with $0 \leq d_k \leq n - 1$:

- I A home-break (away-break) at MD 1 is possible if no home-break (away-break) has been set for MD 1 on $\vec{V}_k$.
- II If $nb_{p, \vec{V}_k} = 1$ we can set a home-break (away-break) at MD p if the existing one is an away-break (home-break).
- III We can set a break for a MD $p \geq 2$ with $nb_{p, \vec{V}_k} = 0$ if $|P_{\vec{V}_k}^{br}| < \frac{n}{2} - 1 + |\{1\} \cap P_{\vec{V}_k}^{br}|$.

Rule I states that team $d_k + 1$ might have a break at MD 1 while less than two breaks are arranged at MD 1 on $\vec{V}_k$. This is according to characteristic iii. Rule II takes care of the fact that at each MD either two or no breaks are located according to ii. Hence, if exactly one break is arranged at MD p on $\vec{V}_k$ team $d_k + 1$ might have a break at MD p . Rule III decides whether a break can occur at a MD where no break is arranged on $\vec{V}_k$ yet. Since there can be no more than $\frac{n}{2}$ MDs having breaks (including MD 1) this is not possible if $\frac{n}{2}$ MDs (including MD 1) or $\frac{n}{2} - 1$ MDs (excluding MD 1) already have breaks. Hence, if no break can be arranged at MD p having $nb_{p, P_k} = 0$ all breaks to be set on the way down the branching tree are forced to fill other MDs up to exactly 2 breaks according to rules I and II.

We employ depth first search (DFS) as node order strategy. Children of a particular node k are inserted in ascending order of team d_k 's break's period. Obviously,

each node k might have up to $2n - 2$ children leading to a rather wide search tree. In depth n of the tree $n(\frac{3}{4}n^2 - \frac{7}{4}n + 1)$ of originally $n(n - 1)^2$ match variables are fixed to zero. However, as lined out in [5] characteristics i to iv are necessary but not sufficient for a MBSRRT. Therefore, we can not totally prevent the routine to investigate an infeasible subtree.

We can improve infeasibility detection by incorporating a further necessary condition from [5]: due to restriction (2) for each subset $T' \subset T$ there must be exactly $\frac{|T'|(|T'|-1)}{2}$ matches of teams of T' against each other.

Theorem 1. *A sequence s of three consecutive periods having breaks leads to an infeasible subtree.*

In order to proof theorem 1, first, we introduce home-away-patterns (HAP). A HAP of team i is a string of length $|P|$ having 0 in slot p if i plays at home at MD p and 1 otherwise.

Proof. Let p' be the first period of s . Each period having a break has a home-break and an away-break. We combine three HAPs having breaks in p' , $p' + 1$, and $p' + 2$ such that the break's venue in p' is equal to the break's venue in $p' + 2$ and different from the break's venue in $p' + 1$. The three teams corresponding to these three HAPs cannot play against each other in periods $p'' < p'$ and $p'' > p' + 1$. There can be exactly one match among these teams in period p' and in period $p' + 1$. Therefore, there is a subset of teams which can play only $\frac{|T'|(|T'|-1)}{2} - 1$ times against each other. $\square$

Table 1 provides an example of HAPs combined as done in the proof. According to theorem 1 we modify rule III to rule III':

III' We can set a break for a MD $p \geq 2$ with $nb_{p, \vec{v}_k} = 0$ if $\text{III} \wedge$

$$\left(\left(nb_{p-2, \vec{v}_k} = 0 \vee nb_{p-1, \vec{v}_k} = 0 \right) \wedge \left(nb_{p-1, \vec{v}_k} = 0 \vee nb_{p+1, \vec{v}_k} = 0 \right) \wedge \left(nb_{p+1, \vec{v}_k} = 0 \vee nb_{p+2, \vec{v}_k} = 0 \right) \right) .$$

Rule III' checks whether a sequence of three periods having breaks would be arranged if a break is set in period p .

Table 1. Example for 3 HAPs with too few matches

period	1	...	$p' - 1$	p'	$p' + 1$	$p' + 2$	$p' + 3$	...	$ P $
HAP 1	...	...	0	0	1	0	1	...	...
HAP 2	...	...	0	1	1	0	1	...	...
HAP 3	...	...	0	1	0	0	1	...	...

Since not each subtree can be pruned at depth n (not even if the subtree is feasible) we have to decide how branching is done at deeper nodes. In order to line out the advantage of our branching idea for guiding the first branches we employ standard mechanisms to solve the subproblem corresponding to a node k with $d_k = n$. Therefore, the 0-1 program corresponding to node k is solved to optimality using CPLEX 9.0.

4 Computational Results

We solved ten instances each for $n = 6$ and $n = 8$ and three instances having $n = 10$, respectively, where $c_{i,j,p}$ has been chosen at random from $[-10, 10]$. The tests were executed on a 3.8 GHz Pentium IV PC with 3 GB RAM running Windows Server 2003. Computation times are provided in table 2.

Table 2. Running times

size	CPLEX 9.0 DFS	CPLEX 9.0 BFS	HAP Branch DFS
6	0.2283 sec	0.2126 sec	0.6079 sec
8	88.113 sec	42.678 sec	91.531 sec
10	not optimal/feasible within 6 days	out of memory	6.82 hours

Using CPLEX with best first search (BFS; that is, exploring the node having the best bound first) outperforms DFS and our branch and bound approach for instances having up to 8 teams. For 10 teams, however, BFS runs out of memory. To the contrary, DFS has lower memory requirements but the instance can not be solved within 6 days. In fact, not even a single feasible solution has been found using DFS whereas our algorithm terminates with a provably optimal solution within 9.5 hours. Moreover, it takes only 9.7 seconds on average to compute a feasible solution.

5 Conclusions and Outlook

The problem to schedule a sports league with a minimum number of breaks based on matches' cost turned out to be intractable using a standard IP solver for more than 8 teams. We propose a branching idea based on properties of HAP sets having a minimum number of breaks. The proposal proofed to be quiet effective in cutting infeasible subtrees of a branch and bound search tree. This enables us to solve instance with 10 teams within 10 hours to proven optimality. Furthermore, our algorithm generates feasible solutions very fast. Nevertheless, some interesting extensions and variants do exist:

- Dynamic cost oriented branching as well as different node order strategies could accelerate the existing approach.
- Generalizing rule III' towards checking of sequences of more than 3 periods having breaks could lead to cutting more infeasible subtrees as done so far.
- SRRTs providing exactly one break per team are closely related to those having a minimum number of breaks. Therefore, our approach should be adaptable to the former case.
- Forbidding breaks in the second period is a common restriction from real world leagues. Adaption of our approach to this restriction is possible.

References

1. T. Bartsch, A. Drexl, and S. Kröger (2006) Scheduling the Professional Soccer Leagues of Austria and Germany. *Computers & Operations Research*, 33: 1907 - 1937
2. D. Briskorn, A. Drexl, and F. C. R. Spieksma (2006) Round Robin Tournaments and Three Index Assignment. Working Paper
3. D. de Werra. *Geography, Games and Graphs* (1980) *Discrete Applied Mathematics*, 2: 327-337
4. F. Della Croce and D. Oliveri (2006) Scheduling the Italian Football League: An ILP-Approach. *Computers & Operations Research*, 33:1963-1974
5. R. Miyashiro, H. Iwasaki, and T. Matsui (2003) Characterizing Feasible Pattern Sets with a Minimum Number of Breaks. In: E. Burke and P. de Causmaecker (eds) *Proceedings of the 4th international conference on the practice and theory of automated timetabling*. *Lecture Notes in Computer Science* 2740, pages 78-99. Springer, Berlin, Germany, 2003.

Priority-Rule Methods for Project Scheduling with Work Content Constraints

Cord-Ulrich Fündeling

Departement Betriebswirtschaftslehre, Universität Bern, Switzerland
fuendeling@ifm.unibe.ch

Summary. In many practical applications of resource-constrained project scheduling a work content (e.g., in man months) is specified for each activity instead of a fixed duration and fixed resource requirements. In this case, the resource usages of the activities may vary over time. The problem then consists in determining the resource usage of each activity at each point in time such that precedence constraints, resource scarcity, and specific constraints on the evolution over time of resource usages are respected and the project duration is minimized. We present two priority-rule methods for this problem and report on results of an experimental performance analysis for instances with up to 200 activities.

1 Introduction

The resource-constrained project scheduling problem $PS|prec|C_{max}$ consists in determining a start time for each project activity such that precedence and resource constraints are met and the project duration is minimized. It is usually assumed that the activity durations as well as the resource requirements of the activities are fixed. For a review of resource-constrained project scheduling we refer to [5,1,10].

For certain practical applications, however, it is more convenient to specify a work content in resource-time units, e.g., in man-hours, for each activity that has to be accomplished. These applications include aggregate project scheduling where detailed activity information may be unknown or uncertain as well as scheduling of labor-intensive projects, e.g., in research and development. The resource usages of an activity may then vary during its execution. Thereby, specific constraints on the evolution over time of resource usages have to be taken into account (work content constraints). The duration of an activity results implicitly from its evolution of resource usages.

This modification of $PS|prec|C_{max}$ which will be referred to as the project scheduling problem with work content constraints has been treated by [3,2,6,7], as well as [8]. None of these approaches seems to be capable of solving problem instances of practical size taking into account all the work content constraints considered in

this paper. We therefore present two priority-rule methods. The first priority-rule method is based on a serial generation scheme that simultaneously determines the evolution of resource usages and the start time for an activity at a time. The second priority-rule method makes use of a parallel generation scheme which schedules parts of the activities in parallel. To this end, certain decision periods are considered successively. For details on both methods, we refer to [4].

2 Problem Statement

We assume that n non-interruptible project activities $1, \dots, n$ have to be scheduled. The set of all activities is denoted by V . Between two activities $i, j \in V$ ($i \neq j$), there may be a *precedence constraint* $\langle i, j \rangle$ indicating that activity j may start as soon as activity i has been completed. In this case, i (j) is called a *direct predecessor* (*direct successor*) of j (i). If there are precedence constraints $\langle i, \iota_1 \rangle, \langle \iota_1, \iota_2 \rangle, \dots, \langle \iota_m, j \rangle$ with $\iota_1, \iota_2, \dots, \iota_m \in V$ ($m > 0$), i (j) is called a *indirect predecessor* (*indirect successor*) of j (i).

Several *renewable resources* with limited capacity are needed for the execution of the activities. For each activity $i \in V$, a *work content* is given that has to be processed by one of these renewable resources. This resource will be referred to as the activity-independent *work content resource*. At each point in time during execution of activity i , the number of units of the work content resource required to process i may vary between a given *minimum* and a given *maximum resource usage*. Furthermore, the length of some time interval during which no variation of the usage of work content resource by some activity $i \in V$ takes place may not be smaller than a given number of time units which will be called the *minimum block length*.

With respect to the renewable resources different from the work content resource, we assume that their usage by some activity $i \in V$ is a non-decreasing function of the respective usage of the work content resource. Eventually, we assume that the resource usages of each activity can only be changed at integral points in time.

The *project scheduling problem with work content constraints* then consists in determining an *activity profile*, i.e., an assignment of a usage of the work content resource to each time period, for each activity $i \in V$ such that

- (1) the precedence constraints are met,
- (2) the resource capacity of neither the work content resource nor any other resource is exceeded during any period,
- (3) for each activity, the number of resource-time units processed without interruption equals its work content,
- (4) the resource usage of each activity does not fall below the activity's minimum resource usage in any period where the activity is processed,
- (5) the resource usage of each activity does not exceed the activity's maximum resource usage in any period where the activity is processed,
- (6) for each activity, the minimum number of successive periods without change of the resource usage is not less than the given minimum block length,

and the project duration is minimized. Constraints (3) to (6) are referred to as the *work content constraints*.

3 Solution Methods

3.1 Serial Generation Scheme

The serial generation scheme is visualized by Figure 1. Set $\mathcal{C}$ comprises the activities that have already been scheduled. As long as $\mathcal{C} \neq V$ holds iterations of the serial generation scheme are performed. In each iteration, we choose an activity $i' \in V \setminus \mathcal{C}$ with highest priority value among all the *eligible* activities whose predecessors have already been scheduled. This activity i' is then started as early as possible and processed as fast as possible. To this end, we first try to start i' in its earliest start period $ES_{i'}$ which is computed taking into account minimum durations for all the activities. If no feasible activity profile for i' is found due to the violation of some resource constraint or work content constraint (3) the start of i' is iteratively delayed by one period. Finally, activity i' is added to set $\mathcal{C}$.

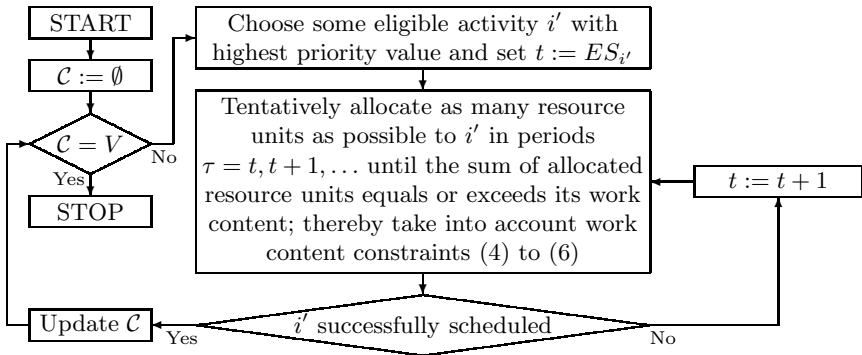


Fig. 1. Serial generation scheme

3.2 Parallel Generation Scheme

The proceeding of the parallel generation scheme is illustrated by Figure 2. In each iteration of the generation scheme some decision period t is considered. For this decision period we first ensure that *active* activities who have been started in an earlier period but have not yet been completed are continued. In case that for some active activity the minimum block length has already been reached by period t we determine a minimum feasible resource usage for this activity. This minimum feasible resource usage ensures that the activity can be completed satisfying all the work content constraints. It may be greater than the resource usage that has been chosen for the activity in the preceding period. Therefore, a resource conflict may occur which is resolved by some unscheduling step that leads to a decrease of the resource usages chosen in the preceding periods.

Moreover, set $\mathcal{Z}$ of activities whose resource usage may be modified in decision period t is established. This set contains all activities for which the minimum block length has already been reached by period t for the last chosen resource usage as well

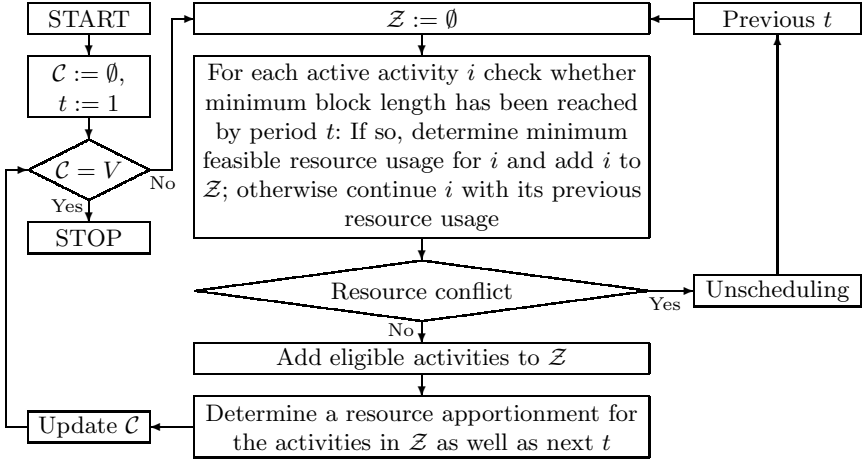


Fig. 2. Parallel generation scheme

as the *eligible* activities which are defined to be those activities whose predecessors have already been completely scheduled.

Finally, we determine an apportionment of the capacity of the work content resource to the activities from set Z . Thereby, we take into account the minimum feasible resource usages as well as the priority values of the activities. Furthermore, the next decision period t is determined as the period in which a modification of the resource usage of some activity from Z will be necessary in order to satisfy the work content constraints. Finally, all activities for which the total work content will have been processed by this next decision period are added to set C which comprises all activities that have already been completely scheduled.

4 Computational Results

For evaluating the priority-rule methods, we have generated a set of 2400 instances by a full factorial experiment on the following parameters: number of activities, order strength, resource factor, and resource strength (cf. [9]). The instances contain up to 200 activities and 4 resources each. The experimental performance analysis has been performed on an Intel Pentium IV PC with 3.0 GHz clockpulse and 1 GB RAM. We have used an adapted version of the Branch-and-Bound algorithm of Demeulemeester et al. [2] as a benchmark procedure (BM). This algorithm as well as both priority-rule methods have been implemented in ANSI-C using Microsoft Visual C++ 6.0.

Single-start as well as multi-start experiments have been performed for both priority-rule methods. In each single-start experiment, one of the following priority rules has been applied:

- **MWR-rule:** The priority value of some activity $i \in V$ corresponds to the sum of its work content and the work contents of all its direct and indirect successors.

- **LPF-rule:** The priority value of i corresponds to a lower bound on the number of time periods needed to process i and all its direct and indirect successors.
- **MTS-rule:** The priority value of i corresponds to the number of all direct and indirect successors of i .

For the multi-start experiments, a time limit of 30s has been prescribed during which the priority-rule methods have been restarted as often as possible using random priority values. For the benchmark procedure, the time limit has been set to 30s as well.

Computational results are given by Table 1. Both priority-rule methods find a feasible solution for every instance (cf. p_{feas}). Only 18.6 % of the instances can be solved to feasibility by the benchmark procedure. None of these instances contains more than 20 activities. For the instances solved to feasibility by the benchmark procedure, however, the quality of solutions in terms of p_{opt} (percentage of instances solved to optimality) and Δ_{LB} (average relative deviation from an appropriate lower bound) is rather good. Computational times for finding the first (t_{first}^{CPU}) as well as the best (t_{best}^{CPU}) solution are small for all algorithms.

Table 1. Computational results

	Serial generation scheme				Parallel generation scheme				BM
	MWR	LPF	MTS	Multi-Start	MWR	LPF	MTS	Multi-Start	
p_{feas} [%]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	18.6
p_{opt} [%]	0.3	0.4	0.3	0.5	0.3	0.4	0.4	0.5	5.5
Δ_{LB} [%]	21.5	21.0	21.6	20.5	23.1	22.9	22.9	21.1	8.1
t_{first}^{CPU} [s]	0.4	0.2	0.4	0.2	0.6	0.3	0.6	0.3	2.0
t_{best}^{CPU} [s]				4.7				6.1	2.0

5 Conclusions and Outlook

In this contribution, we have presented two priority-rule methods for a modification of $PS|prec|C_{max}$ where a work content is given for each activity instead of a fixed duration and fixed resource requirements. One method is based on a serial generation scheme. The other one makes use of a parallel generation scheme where each activity is scheduled in one or more successive iterations. Computational results show that both methods are capable of solving large problem instances with up to 200 activities.

Interesting directions of future research include the consideration of several variations of the project scheduling problem with work content constraints. For instance, progress-dependent precedence constraints should be considered. Moreover, different financial or resource-based objective functions, which may be used for aggregate project scheduling, should be taken into account. Furthermore, the generation schemes should be integrated into evolutionary search strategies in order to further improve the solution quality.

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. Demeulemeester E, De Reyck B, Herroelen W (2000) The discrete time/resource trade-off problem in project networks: A branch-and-bound approach. *IIE Transactions* 32:1059–1069
3. De Reyck B, Demeulemeester E, Herroelen W (1998) Local search methods for the discrete time/resource trade-off problem in project networks. *Naval Research Logistics* 45:553–578
4. Fündeling CU (2006) Ressourcenbeschränkte Projektplanung bei vorgegebenen Arbeitsvolumina. Deutscher Universitäts-Verlag, Wiesbaden
5. Herroelen W, De Reyck B, Demeulemeester E (1998) Resource-constrained project scheduling: a survey of recent developments. *Computers and Operations Research* 25:279–302
6. Kolisch R, Meyer K, Mohr R, Schwindt C, Urmann M (2003) Ablaufplanung für die Leitstrukturoptimierung in der Pharmaforschung. *Zeitschrift für Betriebswirtschaft* 73:825–848
7. Kuhlmann A (2003) Entwicklung eines praxisnahen Project Scheduling Ansatzes auf der Basis von Genetischen Algorithmen. Logos, Berlin
8. Meyer K (2003) Wertorientiertes Projektmanagement in der Pharmaforschung. Shaker, Aachen
9. Neumann K, Schwindt C, Zimmermann J (2003) Project Scheduling with Time Windows and Scarce Resources. Springer, Berlin
10. Neumann K, Schwindt C, Zimmermann J (2006) Resource-constrained project scheduling with time windows: Recent developments and new applications. In: Weglarz J, Jozefowska J (eds) *Topics in modern project scheduling*. Kluwer, Boston

Entscheidungsunterstützung für die Projektportfolioplanung mit mehrfacher Zielsetzung

Antonia Maria Knübel and Natalia Klierwer

Decision Support and Operations Research Laboratory, Universität Paderborn,
Germany
knuebel@dsor.de
klierwer@dsor.de

Summary. Die Projektportfolioplanung stellt ein multikriterielles Entscheidungsproblem dar, das neben der Projektauswahl unter Berücksichtigung von Projektabhängigkeiten und Synergieeffekten verschiedener Art auch die zeitliche Einplanung von Projekten umfasst. Dabei sollen auch die Begrenzungen für den Ressourcenverbrauch von erneuerbaren und nicht erneuerbaren Ressourcen beachtet werden. Um eine Entscheidungsunterstützung hierfür zu bieten, wird zunächst ein mathematisches Modell aufgestellt, um darauf folgend den Einsatz von Standardoptimierungssoftware wie auch von heuristischen Verfahren als Möglichkeit zum Umgang mit mehreren Zielfunktionen vorzustellen.

1 Einleitung

In der Projektportfolioplanung, der zentralen Aufgabe des Multiprojektmanagements, wird die Entscheidung über die konkret durchzuführenden Projekte und damit die Zusammensetzung der Projektlandschaft eines Unternehmens getroffen. Dazu werden in der Projektportfolioplanung, wie sie für diese Arbeit verstanden wird, ausschließlich unternehmensinterne Projekte betrachtet, wobei zunächst keine Unterscheidung nach Projektarten vorgenommen wird. Praktische Relevanz erhält die Projektportfolioplanung, weil durch eine hohe Anzahl interner Projekte in vielen Unternehmen ein beträchtliches Maß an Ressourcen gebunden wird. Das übergeordnete Ziel der Projektportfolioplanung ist die Nutzenmaximierung des Unternehmensprojektportfolios, welches impliziert, dass das Portfolio in seiner Gesamtheit betrachtet wird und keine Beschränkung ausschließlich auf die Beurteilung auf Ebene der Einzelprojekte stattfindet. Kriterien für die Einplanung eines Projektes in das Projektportfolio eines Unternehmen sind vielfältig: Strategiebeitrag, Risiken, Ressourcenverbrauch von erneuerbaren und nicht erneuerbaren Ressourcen, zeitliche Projektanordnung, Wirtschaftlichkeit, operative Dringlichkeit und logische Abhängigkeiten verschiedener Art zwischen einzelnen Projekten. Diese Faktoren dienen als

Beurteilungsmaßstab für Einzelprojekte sowie für den Gesamtnutzen eines Portfolios und zeigen ein Entscheidungsproblem mit mehreren Zielsetzungen und vielfältigen Restriktionen auf.

Für die Lösung des Entscheidungsproblems sind verschiedene Ansätze bekannt, die im Rahmen einer Verwendung von OR-Methoden meist nur Teile der Problemstellung betrachten oder sich auf Projektportfolios beziehen, welche nur eine bestimmte Art von Projekten enthalten. In Bezug auf die verwendeten Methoden lassen sich jene Ansätze insofern kommentieren, als dass sie den Umgang mit mehreren Zielfunktionen in den häufigsten Fällen durch eine Aggregation dieser zu einer einzigen Zielfunktion lösen und vernachlässigen, dass bei mehrfachen Zielsetzungen durchaus verschiedene nicht dominierte Projektportfolios aus einem Pool von Projekten ermittelt werden können.

2 Problemdefinition

Die Grundlage für das folgende mathematische Modell bildet der Ansatz von Archer et al. [2], der versucht mit Hilfe einer aggregierten, zu maximierenden Zielfunktion unter Beachtung einer Reihe von Restriktionen ein optimales Projektportfolio zu finden. Die Restriktionen (2), (3), (4), (5) und (6) sind diesem Ansatz entnommen; da das Modell aber für die oben beschriebene Problemstellung nicht ausreichend ist, wird es um die übrigen Restriktionen erweitert. Der hier vorgestellte Ansatz zeichnet sich im Gegensatz zu [2] dadurch aus, dass bei Beibehaltung aller Zielfunktionen die wesentlichen Restriktionen berücksichtigt und zudem Projekte zeitlich innerhalb des vorgegebenen Planungshorizontes eingeplant werden.

Definitionen

I. Mengen

M_{vor}^l : Menge der Projekte, die für Projekt l Voraussetzung sind

M_{syn}^s : Menge der Projekte, die zusammen eine Synergie s ergeben

M_{max}^w : Menge der Projekte, die in der Gruppe w enthalten sind

M_{muss} : Menge der Mussprojekte

M_{nach} : Menge der Projekte, die einen oder mehrere Vorgänger $i \in M_{vor}^l$ benötigen

M_{zus}^h : Menge der Projekte in Gruppe h , die entweder zusammen oder gar nicht ausgeführt werden dürfen

II. Parameter

N : Anzahl der Projekte

T : Anzahl der Perioden, in die der Planungszeitraum unterteilt wird

U : Anzahl der Zielfunktionen

α_i^u : Beitrag, den Projekt i an der Zielfunktion u liefert

V : Anzahl der erneuerbaren Ressourcen

$C_{i,k+1-j}^v$: erneuerbare Ressource v , die Projekt i in der Periode k verbraucht, wenn es in der Periode j gestartet wird

RL_k^v : Ressourcenlimit einer erneuerbaren Ressource v in der Periode k

- Q : Anzahl der nicht erneuerbaren Ressourcen
 CG_i^q : Gesamtbedarf von Projekt i an der nicht erneuerbaren Ressource q
 RLG^q : Ressourcenlimit von nicht erneuerbarer Ressource q im Planungszeitraum
 D_i : Dauer des Projekts i
 L : Anzahl der Projekte, die einen Vorgänger besitzen
 S : Anzahl der Synergien
 g_s : Anzahl der Projekte, die zusammen eine Synergie s ergeben
 b_s^u : Effekt, den Synergie s auf die Zielfunktion u hat
 W : Anzahl der Gruppen, aus der nur eine maximale Zahl an Projekten ausgeführt werden dürfen
 R^w : Maximale Anzahl von Projekten einer Gruppe w , die im Portfolio enthalten sein dürfen
 H : Anzahl der Gruppen, die Projekte enthalten, die entweder zusammen oder gar nicht ausgeführt werden dürfen
 FS_i : Frühester Startzeitpunkt für Projekt i
 SE_i : Spätester Startzeitpunkt für Projekt i

III. Variablen

- x_{ij} : Binäre Variable, die 1 ist, falls Projekt i in der Periode j gestartet wird
 y_s : Binäre Variable, die 1 ist, falls die Synergie s besteht
 f^h : Binäre Variable, die 1 ist, falls alle Projekte einer Gruppe h gleichzeitig ausgeführt werden

Modell

I. Zielfunktionen

$$\min/\max \sum_{i=1}^N \sum_{j=1}^T a_i^u x_{ij} + \sum_{s=1}^S b_s^u y_s \quad u = 1, \dots, U \quad (1)$$

In den Zielfunktionen ist der Effekt angegeben, den ein Projekt i auf die jeweilige Zielfunktion u hat. Falls nun das Projekt i im Portfolio enthalten ist, wirkt sich dies durch den Zielfunktionskoeffizienten aus. Außerdem sind die Auswirkungen modelliert, die auftreten, falls alle Projekte einer Synergie s durchgeführt werden sollen und damit ein Synergieeffekt für eine oder mehrere Zielfunktionen auftritt (s. a. [3]).

II. Restriktionen

$$\sum_{j=1}^T x_{ij} \leq 1 \quad \forall \quad i = 1, \dots, N \quad (2)$$

$$\sum_{j=1}^T x_{ij} = 1 \quad \forall \quad i \in M_{muss} \quad (3)$$

$$\sum_{j=1}^T x_{ij} \geq \sum_{j=1}^T x_{lj} \quad \forall \quad l \in M_{nach}, i \in M_{vor}^l \quad (4)$$

$$\sum_{j=1}^T j \cdot x_{lj} + (T+1)(1 - \sum_{j=1}^T x_{lj}) - \sum_{j=1}^T j \cdot x_{ij} \geq D_i \sum_{j=1}^T x_{ij} \quad \forall \quad l \in M_{nach}, i \in M_{vor}^l \quad (5)$$

$$\sum_{i=1}^N \sum_{j=1}^k C_{i,k+1-j}^v x_{ij} \leq RL_k^v \quad \forall \quad k = 1, \dots, T, v = 1, \dots, V \quad (6)$$

$$\sum_{i=1}^N \sum_{j=1}^T CG_i^q x_{ij} \leq RLG^q \quad \forall \quad q = 1, \dots, Q \quad (7)$$

$$\sum_{i \in M_{syn}^s} \sum_{j=1}^T x_{ij} \geq g_s y_s \quad \forall \quad s = 1, \dots, S \quad (8)$$

$$g_s - \sum_{i \in M_{syn}^s} \sum_{j=1}^T x_{ij} \geq 1 - y_s \quad \forall \quad s = 1, \dots, S \quad (9)$$

$$\sum_{i \in M_{max}^w} \sum_{j=1}^T x_{ij} \leq R^w \quad \forall \quad w = 1, \dots, W \quad (10)$$

$$\sum_{j=1}^T x_{ij} = f^h \quad \forall \quad h = 1, \dots, H, i \in M_{zus}^h \quad (11)$$

$$\sum_{j=1}^{FS_i-1} x_{ij} = 0 \quad \forall \quad i = 1, \dots, N \quad (12)$$

$$\sum_{j=SE_i+2-D_i}^T x_{ij} = 0 \quad \forall \quad i = 1, \dots, N \quad (13)$$

Restriktionen (2) sichern, dass jedes Projekt nur genau einmal oder nicht gestartet wird. Den Start von Mussprojekten für einen Startzeitpunkt im Planungszeitraum gewährleisten Restriktionen (3). In den Restriktionen (4) und (5) werden die Vorgänger-Nachfolger-Beziehungen zwischen Projekten modelliert. Restriktionen (4) sorgen dafür, dass das Projekt l nicht im Projekt-

portfolio enthalten ist, falls die Vorgänger i nicht durchgeführt werden. Restriktionen (5) gewährleisten, dass die von Projekt l benötigten Vorgänger i auch wirklich beendet wurden, wenn Projekt l gestartet wird. Mit den Restriktionen (6) und (7) wird sichergestellt, dass durch die eingeplanten Projekte nicht mehr Ressourcen verbraucht werden, als tatsächlich zur Verfügung stehen. Beim Auftreten von Synergieeffekten gewährleisten die Restriktionen (8) und (9), dass eine entsprechende Änderung eines oder mehrerer Zielfunktionswerte erfolgt, falls eine bestimmte Gruppe von Projekten zum Portfolio gehört. Mit Hilfe von (10) wird garantiert, dass aus einer Gruppe von Projekten nur eine bestimmte Anzahl maximal im gleichen Portfolio enthalten ist. Die Restriktionen (11) sichern, dass Projekte einer Alle-oder-keine-Gruppe ausnahmslos den gleichen Status, entweder durchführen oder nicht durchführen, erhalten. Schließlich werden durch die Restriktionen (12) und (13) noch die Variablenwerte für Projekt i zu den Zeitpunkten, die vor einem frühesten Startzeitpunkt und nach einem spätesten Endzeitpunkt liegen, auf Null fixiert. Zu den Teilaspekten der in diesem Abschnitt vorgenommenen Charakterisierung der Entscheidungssituation vergleiche auch [5, 4, 6, 2].

3 Lösungsverfahren

Für die Lösung des oben beschriebenen Problems kommen verschiedene Verfahren in Betracht; dazu zählen sowohl exakte als auch heuristische Methoden. Im Rahmen dieser Arbeit wurden folgende Verfahren getestet:

- Optimierung mittels Standard-Optimierungssoftware nach jeweils einer Zielfunktion sowie Einsatz der Tchebycheff Metrik (vgl. [1])
- Anwendung des Ameisenkolonie-Algorithmus nach [3] als heuristisches Verfahren zur Mehrziel-Projektportfolioplanung mit dem Ziel, in möglichst kurzer Zeit eine große Menge an pareto-optimalen Projektportfolios zu finden, und Anpassung für das oben beschriebene Problem
- Einsatz der paretobasierten, evolutionären Metaheuristiken NSGA 2 (Non-dominated Sorting Genetic Algorithm 2) und SPEA 2 (Strength Pareto Evolutionary Algorithm 2) [7].

4 Ergebnisse und Fazit

Für die bisherigen Tests der genannten Lösungsverfahren wurden verschiedene fiktive Probleminstanzen verwendet. Dabei variierten die Testinstanzen in der Anzahl der Projekte bei gleichbleibenden restlichen Parametern. Bei den Ergebnissen zeichnet sich innerhalb der Gruppe der metaheuristischen Lösungsverfahren eine deutliche Überlegenheit des Ameisenalgorithmus ab. Bei vorgegebener maximaler Laufzeit von einer Minute fand dieser im Vergleich zu den genetischen Algorithmen SPEA2 und NSGA2 durchschnitt-

lich die doppelte Anzahl nicht-dominiierter Projektportfolios. Die Standard-Optimierungssoftware war aufgrund der bisher vornehmlich kleinen Testinstanzen innerhalb weniger Sekunden in der Lage, die optimalen Lösungen zu ermitteln. Weitere Parametertests der Metaheuristiken werden Aufschluss darüber geben, ob die Überlegenheit des Ameisenalgorithmus im Vergleich zu den genetischen Algorithmen Bestand hat.

Die Anwendung von Heuristiken zur Lösung des oben vorgestellten umfassenden Modells für die Projektportfolioplanung bietet einem Entscheider durch die Ermittlung nicht-dominierter Projektportfolios eine echte Entscheidungshilfe, da dieser zwischen Portfolios auswählen kann. Im Gegensatz dazu ergibt die Verwendung von Standard-Optimierungssoftware oder der Tchebycheff Metrik nur ein einzelnes Portfolio, das aber im Entscheidungsprozess durchaus zum Vergleich mit anderen Projektportfolios dienen kann.

References

1. Tkindt V, Billaut J (2002) Multicriteria Scheduling. Springer, Berlin
2. Archer N, Ghasemzadeh F, Iyogun P (1999) A zero-one model for project portfolio selection and scheduling. *Journal of Operational Research Society*
3. Doerner K, Gutjahr W, Hartl R, Strauss C, Stummer C (2006) Pareto ant colony optimization with ILP preprocessing in multiobjective project portfolio selection. *European Journal of Operational Research*
4. Hiller M (2002) Multiprojektmanagement-Konzept zur Gestaltung, Visualisierung, Regelung und Visualisierung einer Projektlandschaft. FBK Produktionstechnische Berichte der Universität Kaiserslautern
5. Kunz C (2005) Strategisches Multiprojektmanagement-Konzeption, Methoden und Strukturen. Gabler, Wiesbaden
6. Stummer C (1998) Projektauswahl im betrieblichen F&E-Management. Gabler, Wiesbaden
7. Zitzler E, Laumanns M, Künzli S, Bleuler S, Thiele L (2003) <http://www.tik.ee.ethz.ch/pisa/>.

Eine Web-Service basierte Architektur für ein Multi-Agenten System zur dezentralen Multi-Projekt Planung

Jörg Homberger, Raphael Vullriede, Jörn Horstmann, René Lanzl,
Stephan Kistler, and Thomas Göttlich

Fachhochschule Kaiserslautern, Pirmasens, Germany
`joerg.homberger@fh-kl.de`

Summary. Es wird die Architektur eines entwickelten Multi-Agenten Systems (MAS) zur dezentralen Lösung des Decentral Resource Constrained Multi Projekt Scheduling Problem (DRCMPSP) vorgestellt. Die Systemarchitektur basiert auf Web-Services, welche die einfache Integration und somit auch Evaluation alternativer Koordinationsmechanismen von Agenten ermöglichen. Bei dem System handelt es sich um eine internetbasierte Anwendung, deren Funktionalität über einen Internetbrowser zur Verfügung gestellt wird (<http://www.agentcopp.de>). Auf diese Weise kann das MAS auch zukünftig von Systementwicklern und Wissenschaftlern in den von ihnen gewählten IT-Umgebungen benutzt und mit anderen Systemen unter identischen IT-Rahmenbedingungen verglichen werden. Ferner werden 80 Benchmarkprobleme und Lösungen für das DRCMPSP präsentiert. Mit dem Beitrag wird das Ziel verfolgt, das Benchmarking von Systemen für das DRCMPSP transparenter zu gestalten.

1 Problembeschreibung und Zielsetzung

Dem folgenden Beitrag liegt das Decentral Resource Constrained Multi-Projekt Scheduling Problem (DRCMPSP) zugrunde (vgl. [1], [2]). Dieses besteht darin, gleichzeitig mehrere eigenständige (Einzel-)Projekte dezentral zu planen (zur Planung von Multi-Projekten vgl. auch [3]). Hierbei wird jedes einzelne Projekt in Anlehnung an das Resource Constrained Projekt Scheduling Problem (RCPSP) modelliert (vgl. [4]). Bei der Multi-Projekt Planung wird für jedes Projekt zusätzlich eine frühest mögliche Startperiode vorgegeben. Zur Durchführung der Projekte steht eine Menge erneuerbarer Ressourcen zur Verfügung, die von den Projekten geteilt werden. Für jede Ressource und jede Periode des Planungszeitraumes ist eine begrenzte Kapazität gegeben, so dass die Projekte um die knappen Ressourcenkapazitäten konkurrieren und eine Allokation der Ressourcenkapazitäten vorzunehmen ist.

Für jedes Projekt besteht das dezentrale Ziel, den Makespan zu minimieren. Insofern liegt ein multikriterielles Optimierungsproblem vor. Zur Berechnung einer pareto-optimalen Lösung wird ein Lexikographisches MinMax-Fairness-Problem formuliert [5]. Das heisst, eine Multi-Projekt Lösung ist um so fairer, je kleiner der grösste (zweitgrösste etc.) Makespan ist.

In der Literatur werden zur dezentralen Ressourcenallokation MAS empfohlen (vgl. [6], [7], [8])). Bei einem MAS wird die Allokation von Ressourcen durch autonom operierende und miteinander kooperierende Softwareagenten gelöst. Die Kooperation wird dabei über verschiedene Koordinationsmechanismen, z.B. elektronische Verhandlungen und Auktionen, geregelt (vgl. [9]). Für das hier zugrunde gelegte DRCMPSP existieren derzeit nur zwei MAS in der Literatur (vgl. [1] und [2]). Ein Performancevergleich der beiden MAS bzw. ein Vergleich der beiden Systeme mit neu entwickelten Systemen fällt aus folgenden Gründen schwer: Zum einen hängt das Laufzeitverhalten eines MAS massgeblich von der zugrundeliegenden IT-Umgebung, wie z.B. Rechner und Rechnernetze, ab. Zum anderen wurden die MAS von [1] und [2] auf unterschiedlichen, jeweils von den Autoren entwickelten Instanzen, getestet. Die Instanzen sind nicht veröffentlicht und stehen daher für zukünftige Benchmarkstudien nicht zur Verfügung. Darüberhinaus sind die von den Autoren verwendeten Instanzen als klein zu bezeichnen. So generierten [1] Multi-Projekt Instanzen mit lediglich bis zu 9 Projekten und jeweils bis zu 15 Aktivitäten. Die von [2] erzeugten Instanzen bestehen lediglich aus bis zu 5 Projekten mit bis zu 18 Aktivitäten.

Um zukünftig aussagekräftige Benchmarkstudien von MAS für das DRCMPSP zu ermöglichen, wird nachfolgend ein MAS für das DRCMPSP vorgestellt, dessen Funktionalität jedem Systembenutzer über einen Internetbrowser zur Verfügung gestellt wird (<http://www.agentcopp.de>). Da es sich um eine internetbasierte Anwendung handelt, kann das System von jedem Benutzer in der von ihm gewählten IT-Umgebung ausgeführt werden. Damit kann das Systemverhalten des vorgestellten MAS und das Verhalten anderer Systeme in der gleichen IT-Umgebung unmittelbar miteinander verglichen werden. Es werden 80 Benchmarkinstanzen für das DRCMPSP generiert und gelöst. Hierbei werden insbesondere auch grosse Multi-Projekt Instanzen mit bis zu 20 Projekten und mit jeweils bis zu 120 Aktivitäten erzeugt.

2 Benchmarksystem

2.1 Dezentraler Planungsansatz

Das entwickelte MAS besteht aus mehreren, miteinander kooperierenden Softwareagenten. In Analogie zu [2] wird jedes Einzelprojekt durch einen vom Benutzer zugewiesenen Agenten gelöst, d.h. vom Agenten wird ein zulässiger Schedule berechnet. Jeder Agent ist für die dezentrale und autonome Planung genau eines Einzelprojektes zuständig. Hierzu verfügt jeder Agent über

eine Optimierungs- bzw. Schedulingkomponente. Diese beinhaltet einen Evolutionären Algorithmus (EA) zur Lösung des RCPSP (vgl. [10]).

Die Kooperation der Agenten wird über einen elektronischen Marktplatz koordiniert, über den die Agenten gemeinsam die Verteilung der Ressourcen berechnen. Hierzu tauschen die Agenten ausgehend von einer auf dem Marktplatz berechneten Startallokation iterativ Ressourcenkapazitäten untereinander aus. In jeder Iteration berechnen die Agenten alternative Angebote und Nachfragen an Ressourcenkapazitäten und stellen diese auf dem Marktplatz ein. Auf den Marktplatz werden am Ende einer Iteration Angebote und Nachfragen derart zueinander geführt, so dass die über alle Projekte kumulierte Durchlaufzeit eine Reduktion erfährt.

2.2 Architektur

Um ein MAS zu entwickeln, dessen Funktionalität einfach erweitert und über das Internet benutzt werden kann, wurde eine auf Web-Services basierte Architektur gewählt. Web-Services stellen die State-of-the-Art Technologie zur Integration elektronischer Verhandlungen in elektronische Marktplätze dar (vgl. [11]).

Bei der Ausgestaltung der internetgestützten Kommunikation zwischen den Agenten einerseits und einem Marktplatz andererseits wurde der praxisrelevante Fall angenommen, dass die Agenten in eigenständigen, projektplanenden Unternehmenseinheiten oder gar in verschiedenen Unternehmen eingesetzt werden. Der damit verbundene unternehmensübergreifende Einsatz der Anwendung stellt besondere Anforderungen an die Kommunikationstechnologie. Zum einen musste darauf geachtet werden, dass die Kommunikation netzwerk-, plattform- und programmiersprachenunabhängig ist, da nicht davon ausgegangen werden kann, dass alle teilnehmenden Akteure die gleiche technologische Basis benutzen. Zum anderen sollte eine reibungslose Kommunikation über Sicherheitshardware und -software, wie z.B. Firewalls, hinweg gewährleistet werden. Dies erfordert ein Datenübertragungsprotokoll, das von den restriktiven Sicherheitsregeln der Unternehmen nicht blockiert wird, wie z.B. das Hypertext Transfer Protocol (HTTP). Zur Realisierung der Kommunikation zwischen Agenten und Marktplatz wurde auf Simple Object Access Protocol (SOAP) zurückgegriffen. SOAP ist ein vom W3C spezifiziertes, XML-basiertes Protokoll zur Entwicklung von Web-Services. Diese werden über das Internet bereitgestellt und können aufgrund der zugehörigen Dienstbeschreibung in WSDL (Web-Service Description Language) von verschiedensten Plattformen aus genutzt werden. Zur Konstruktion der Web-Services wird als SOAP-Engine die Open Source Implementierung Apache AXIS verwendet. In der Abbildung 1 ist die Architektur des entwickelten MAS dargestellt.

Die Kommunikation zwischen Agenten und Marktplätzen wird durch ein Client-Server Modell realisiert. Die Anwendungsfunktionen des entwickelten MAS, wie die Administration von Marktplätzen und die Realisierung der

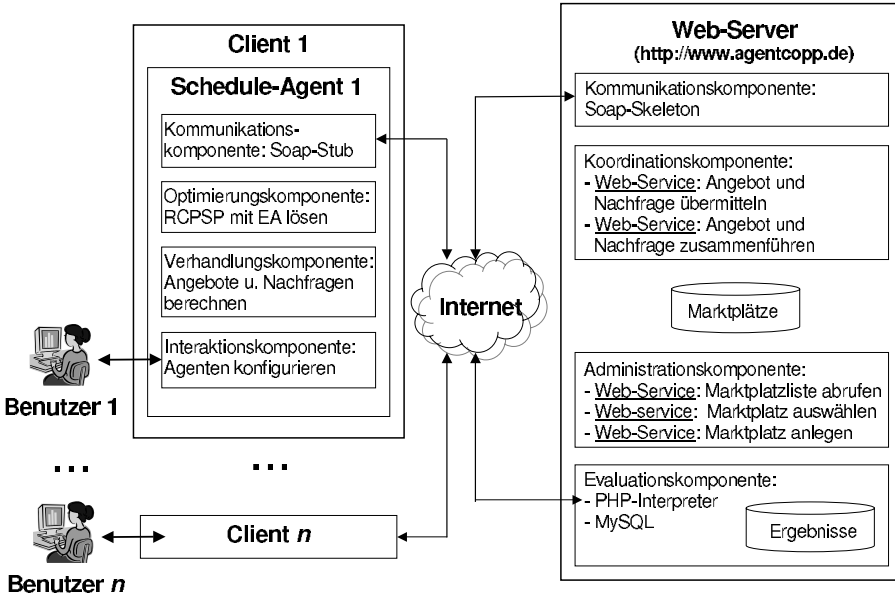


Fig. 1. Architektur

agentenbasierten Verhandlungen, werden durch Web-Services realisiert. Als Beispiel sei der Web-Service "Angebot und Nachfrage übermitteln" erläutert: Hier werden Angebote und Nachfragen von den Agenten auf dem Marktplatz eingestellt. Die Agenten erhalten als Rückgabewerte die auf dem Marktplatz berechneten Ressourcenallokationen. Dieser Web-Service erlaubt eine elegante Kapselung des Koordinationsmechanismus auf dem Server. Auf diese Weise wird eine einfache Erweiterung des Systems um neue Koordinationsmechanismen ermöglicht.

Unter Anwendung der konstruierten Web-Services läuft nun die Lösung des DRCPSP in zwei Phasen ab: (1) Marktplatzadministration und (2) Ressourcenverhandlung. In der ersten Phase können die Systembenutzer über einen gestarteten Agenten eine Liste bereits vorhandener, und auf dem Server abgelegten Marktplätze abrufen und bei Bedarf neue Marktplätze anlegen. Beide Funktionen werden über entsprechende Web-Services realisiert. Anschliessend erfolgt in der zweiten Phase die ebenfalls über Web-Services realisierte und zugleich synchronisierte Verhandlung der Agenten. Zu Beginn wird von den Agenten jeweils über den Web-Service "Marktplatz auswählen" ein Marktplatz selektiert und eine Verbindung zwischen Agent einerseits und dem Marktplatz andererseits aufgebaut. Sobald die vorgegebene Anzahl an Agenten mit dem Marktplatz verbunden ist, erhalten alle Agenten gleichzeitig einen Web-Service Response. Dieser enthält eine Startallokation und wird von den Agenten als Signal interpretiert, mit der iterativen Verhandlung zu beginnen. In jeder Iteration bzw. Verhandlungsrunde berechnen die Agenten

zunächst Angebote und Nachfragen an Kapazitäten. Anschliessend übertragen die Agenten jeweils das lokal berechnete Ergebnis, mit Hilfe des Web-Services "Angebot und Nachfrage übermitteln", an den Marktplatz. Erst wenn alle Agenten ihre Angebote und Nachfragen dem Marktplatz mitgeteilt haben, wird auf diesem eine neue Allokation berechnet. Abschliessend wird das Ergebnis der Verhandlungsrunde, d.h. ein durchzuführender Ressourcenaustausch, allen Agenten gleichzeitig durch Responses der Web-Services übermittelt. Die einzelnen Verhandlungsrunden werden mit Hilfe eines Datenbankservers protokolliert. Über verschiedene Reporting Engines werden dem Benutzer zahlreiche Auswertungen für Benchmarkstudien zur Verfügung gestellt.

3 Benchmarkinstanzen

Mit dem Ziel die Lösungsqualität des entwickelten Systems zu bewerten, wurden auf der Basis der Benchmarkprobleme der Project Scheduling Problem Library (vgl. [12]) 4 Gruppen (M30, M60, M90, M120) mit jeweils 20 Instanzen des DRCMPSP generiert. Jede Instanz besteht aus n ($n = 2, 5, 10, 20$) RCPSP Instanzen. Innerhalb einer Gruppe sind Einzelprojekte gleicher Grösse, d.h. mit jeweils gleicher Anzahl m ($m = 30, 60, 90, 120$) an Aktivitäten, zusammengefasst. Die generierten Multi-Projekt Instanzen wurden mit dem entwickelten Multi-Agenten System gelöst. Sowohl die Instanzen als auch die berechneten Lösungen sind unter <http://www.agentcopp.de> einsehbar und stehen somit zu Vergleichszwecken zur Verfügung.

4 Zusammenfassung und Ausblick

Im Beitrag wurde zum einen die auf Web-Services basierte Architektur eines MAS für das DRCMPSP vorgestellt. Die Funktionalität der internet-gestützten Anwendung kann über jeden Internetbrowser ausgeführt werden (<http://www.agentcopp.de>). Auf diese Weise kann das MAS weltweit genutzt und mit anderen Systemen in gleichen IT-Umgebungen verglichen werden. Zum anderen wurden für das DRCMPSP neue Benchmarkprobleme und erzielte Ergebnisse präsentiert. Somit können in zukünftigen Forschungsarbeiten zum DRCMPSP aussagekräftige Benchmarkstudien durchgeführt werden.

In weiteren Arbeiten ist beabsichtigt, verschiedene Schedulingverfahren, Verhandlungsansätze und marktplatzbasierte Lösungsmodelle zu entwickeln und diese mit Hilfe der vorgestellten Architektur umzusetzen und zu evaluieren. Die auf Web-Services gestützte Systemarchitektur unterstützt hierbei eine einfache Systemerweiterung. Das vorgestellte Multi-Agenten System kann somit dauerhaft als Benchmarksystem im Forschungsbereich des DRCMPSP verwendet werden.

References

1. Lee YH, Kumara SR, Chatterjee K (2003) Multiagent based dynamic resource scheduling for distributed multiple projects using a market mechanism. *Journal of Intelligent Manufacturing* 14:471–484
2. Confessore G, Rismondo S, Giordani S (2005) A market-based multi-agent system model for decentralized multi-project scheduling. *Annals of Operations Research* (als Publikation akzeptiert)
3. Drexel A (1991) Scheduling of project networks by job assignment. *Management Science* 37:1590–1602
4. Brucker P, Drexel A, Moehring R, Neumann K, Pesch E (1999) Resource-constrained project scheduling: notation, classification, models and methods. *European Journal of Operational Research* 112:3–41
5. Ehrgott M (1998) Discrete decision problems, multiple criteria optimization classes and lexicographic max-ordering. In: Stewart TJ, van den Honert RC (Hrsg.) *Trends in multicriteria decision making*. Springer, Berlin
6. Wellman MP, Walsh WE, Wurman PR, MacKie-Mason JK (2001) Auction protocols for decentralized scheduling. *Games and Economic Behavior* 35:271–303
7. Stummer C, Vetschera R (2003) Decentralized planning for multiobjective resource allocation and project selection. *Central European Journal of Operations Research* 11:253–279
8. Chevalere Y, Dunne PE, Endriss U, Lang J, Lemaître M, Maudet N, Padget J, Phelps S, Rodríguez-Aguilar JA, Sousa P (2006) Issues in multiagent resource allocation. *Informatica* (als Publikation akzeptiert)
9. Sandholm TW (1999) Distributed rational decision making. In: Weiss G (Hrsg.) *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT Press, Cambridge MA
10. Hartmann S (2002) A self-adaptive genetic algorithm for project scheduling under resource constraints. *Naval Research Logistics* 49:433–448
11. Rebenstock M, Lipp M (2003). Webservices zur Integration interaktiver elektronischer Verhandlungen in elektronische Marktplätze. *Wirtschaftsinformatik* 45:293–306
12. Kolisch R, Sprecher A (1997) PSPLIB - A project scheduling library. *European Journal of Operational Research* 96:205–216

Approaches to Solving RCPSP Using Relaxed Problem with Consumable Resources

Ivan A. Rykov

Novosibirsk State University, Russia
rykov@ngs.ru

Summary. In the work a resource-constrained project scheduling problem (RCPSP) is considered. This classical NP-hard problem evokes interest from both theoretical and applied points of view. Thus we can divide effective (i.e. polynomial) approximation algorithms in two groups: fast heuristic algorithms for the whole problem and generalizations and algorithms with performance guarantee for particular cases. In first section we consider the statement of the problem and some generalizations. Along with renewable resources we consider consumable resources which can be stored to be used at any moment after the moment of allocation. A polynomial optimal algorithm for solving the problem with consumable resources only was suggested by Gimadi, Sevastianov and Zalyubovsky [2]. So we can consider polynomially solved relaxation of RCPSP. In this relaxation instead of each renewable resource we have consumable resource which is allocated at each moment of time in one and the same amount. Then we can use information about the solution of this relaxation for approximate solving the original problem in polynomial time (for example, the order of starting times can be used as a heuristic for serial scheduling scheme). Furthermore, the optimal value of relaxation gives the lower bound for the optimal value of the original problem.

Speaking about performance guarantee algorithms we assume the case of void precedence graph, and one type of renewable resource. This is a generalization of binary packing problem which looks similar to a strip packing, but not equal to it. We compare this two problems and find bounds for maximal difference of optimal values for them (optimal value of this difference is unknown yet).

1 Problem Definition

1.1 Resource Constrained Project Scheduling Problem (RCPSP)

Classical NP-hard optimization problem, RCPSP can be stated as the following mathematical model:

$$\left\{ \begin{array}{l} \max_j (s_j + p_j) \rightarrow \min_{s_j \in \mathbb{Z}^+} \\ s_i + p_i + w_{ij} \leq s_j, \quad \forall (i, j) \in E(G) \\ \sum_{j \in A(t)} r_{kj} \leq R_k, \quad k = 1 \dots K, t = 1, 2, \dots \end{array} \right.$$

where $A(t) = \{j | s_j < t \leq s_j + p_j\}$.

The goal is to define integer-valued starting times (s_j) for a set of N jobs of fixed integer duration p_j tied with constraints of two types — precedence constraints and resource constraints. Chosen schedule (i.e. the set of s_j) should minimize project finishing time (so-called makespan). Constraints of the first type are defined by the weighted oriented graph $G = (J, E)$ with jobs-nodes. The weight of each node is equal to the duration of the job and the weight of arc represents time lag between the end of one job and beginning of another. It is known that *the schedule that satisfies all precedence constraints exists if and only if G doesn't contain contours of positive weight* (weight of contour is the sum of weights of all nodes and arcs in it). We will consider only contourless graphs that surely satisfies this criteria. Constraints of second type are resource constraints. In the problem it is considered a set of (K) renewable resources. It means that each resource is allocated in each moment of time in one and the same amount and unused units cannot be stored (the examples of such resources are humans, machines, rooms and etc.) For each job there is set amounts of each resource needed at one moment of executing (r_{kj}). Constraints say that at each moment of time for each resource the amount of this resource needed by all executed jobs ($A(t)$) should not exceed allocating amount R_k .

Though it has such clear definition problem is known to be NP-hard. However, feasible schedule, if exists (see criteria above) can be obtained in polynomial time. Fast heuristic algorithms like well-known serial and parallel scheduling schemes use this property.

Generalizations

Besides its internal mathematical beauty, RCPSP is very interesting for the applied science. ERP systems designed for project management and resource planning are very popular in large-scale industry. Often additional constraints generalizing the problem are considered:

release dates: for some jobs there can be defined moments r_j of their appearance in project. Thus, additional constraints look like $s_j \geq r_j, \forall j$. They simply can be modeled in terms of base problem with the help of fictive job of zero duration precede other jobs with lags equal to the release dates
deadlines (due dates): similarly, for some jobs there can be defined moments d_j of their obligatory ending. Corresponding constraints are the following: $s_j + p_j \leq d_j, \forall j$. Unlike release date, deadlines severely complicate the problem: the search of feasible schedule in such problem is as hard as search of optimal solution in base problem, i.e. NP-hard.

resource allocation calendars: the amount of resource allocated at the moment may vary with time passing. This means that we have $R_k = R_k(t), t = 1, 2, \dots$. Also, job during its executing also may require differing amount of resource: $r_{kj}(\tau), \tau = 1, \dots, p_j$. Thus, resource constraints inequalities will look like

$$\sum_{j \in A(t)} r_{kj}(t - s_j) \leq R_k(t), \forall k, t$$

multiple modes: in this generalization each job can be performed in several modes differing in amount of resources demand and duration. When choosing a schedule we have to define both starting times s_j and mode numbers m_j for each job.

resource pools: some resources may be united in pool which becomes an independent resource. If some job requires pool in some amount, we can use any resources from pool in total amount equals to required. Here defining schedule along with s_j we have to define which resources from the pools are used during execution of jobs.

and some other.

Some of this additional instances (like release dates) can be modelled in terms of main problem. Others (multiple modes, resource pools) can't be modelled, but for them finding of some (not optimal) feasible solution can be done in polynomial time. Moreover, it can be shown that classical heuristic schemes (like serial scheduling scheme) can be extended for this cases. The hardest generalizations from the above list are due dates and variable resource allocation and consumption. For them even the search of feasible solution is NP-hard problem.

Types of Resources

In project scheduling sometimes nonrenewable (or consumable) resources are considered along with renewable. Consumable resources are allocated once (at the beginning of the project) and have sense only in multimodal variant of the problem restricting some of the mode combinations.

In soviet (russian) literature it was also popular to consider the third type of resources - storable resources. As renewable resources they are also allocated at each moment of type (in varying amount $R_k(t)$), but unused at some moment units can be stored to use them at any moment later (the best example for this type of resource is money).

1.2 Problem with Consumable Resources

The definition of this problem is quite similar to the RCPSP definition (with due dates and variable resource allocation and demand generalizations). The

difference is that all resources are storable, that's why we have additional summarizing in resource constraints inequalities:

$$\left\{ \begin{array}{l} \max_j (s_j + p_j) \rightarrow \min_{s_j \in Z^+} \\ s_i + p_i + w_{ij} \leq s_j, \quad \forall (i, j) \in E(G) \\ s_i + p_i \leq d_i, \quad \forall i = 1, \dots, N. \\ \sum_{\tau=1}^t \sum_{j \in A(\tau)} r_{kj}(\tau - s_j) \leq \sum_{\tau=1}^t R_k(\tau), \quad k = 1 \dots K, t = 1, 2 \dots \\ \text{where } A(t) = \{j | s_j < t \leq s_j + p_j\} \end{array} \right.$$

(We forget about release dates since they are modeled by precedence constraints.)

Gimadi, Sevastiyarov and Zalyubovsky worked out polynomial optimal algorithm for solving this problem. It is based on the fact that late schedule gives the most comfort situation in usage of consumable resources.

If T is some moment of time than *T-late schedule* (compare with early schedule in critical path analysis) is a schedule with makespan equals to T where each job starts as late as possible satisfying all precedence constraints and due dates. It's easy to understand that if the early (critical path) schedule satisfies all due dates and has the length of T' , then for each $T \geq T'$ there exists T -late schedule. Another important fact for this problem is that if there exists feasible schedule with makespan T , than T -late schedule is also feasible (feasibility in resource constraints is guaranteed because jobs start could only increase and thus the amounts of resources already allocated also could be only higher).

The idea of algorithm is to find the minimal moment of time T such that T -late schedule will be feasible. This moment of time is searched by dichotomy.

Note that even the presence of deadlines here doesn't prevent polynomial complexity. As it was mentioned, in similar problem with renewable resources even the search of feasible solution is NP-hard.

For the instance of RCPSP problem (R) (may be with deadlines) let's consider an instance with the same data where all the resources are regarded as consumable (C). We can notice that any schedule that's feasible for (R) is also feasible for (C). In other words, such operation gives a relaxation of RCPSP. The optimal solution of (C) gives the lower bound for the solution of (R). This lower bound can be significantly better than lower bound given by critical path method, because it takes in account resource constraints, though not in completely proper way.

The order of jobs in optimal solution of (C) gives us a priority rule for a serial scheduling scheme. In the case without deadlines it coincides with LST-rule ([4]), in the case of deadlines we can use it as heuristic in attempts to find feasible solution.

2 RCPSP and Strip Packing Problem

The particular case of RCPSP with $G = \emptyset, K = 1, p_j = 1 \forall j$ is a famous binary packing problem, which is also NP-hard. But since it's much more simple, a variety of algorithms with guaranteed performance bounds.

Strip packing problem is a natural generalization of binary packing problem. This problem was also well investigated in the field of algorithms with bounds. As an input it is given a strip of fixed width W and infinite length, and a set of rectangles with fixed width and length (w_i, h_i) . All rectangles should be placed in strip without intersections in such a way as to minimize the length of packing (i.e. the right side of the rightmost rectangle).

It can be easily noticed that the input of Strip packing instance has one-to-one correspondence with the input of the case of RCPSP with $G = \emptyset, K = 1$ (the width of strip corresponds to the resource allocation amount, rectangles correspond to jobs: their widths — to resource consumption per unit of time, lengths — to jobs' durations). Every feasible packing can be considered as a feasible schedule — if we take x coordinate of the left side of the rectangle as a start time for the corresponding job. However, the converse is not true: not every feasible schedule corresponds to some feasible packing because we do not need to use the same resources each moment of processing. So it is allowed to cut a rectangle-job in pieces of unit length to put them in consecutive layers of the strip.

So the question is: are these problems equivalent, i.e. are optimal values for each instances with the same data identical? It was shown ([3]) that these problems are NOT equivalent. Here is the least instance of the largest known ratio $(= \frac{5}{4})$:

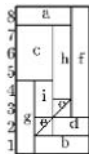


Fig. 1. An example performance ratio $\frac{5}{4}$

On the picture it is shown a feasible schedule with makespan equals to 4. Since it's equal to the lower bound (summary area of jobs divided by resource allocation amount) we can say that it's an optimal value. At the same time, it can be shown that there doesn't exist any feasible packing of length 4. And packing of length 5 is easily constructed and optimal. Thus, we get

$$\frac{SP(I)}{RCPS(I)} = \frac{5}{4}$$

for this instance. Let's consider value

$$\rho = \sup_I \frac{SP(I)}{RCPS(I)}$$

where I runs through all the instances of the problems. The above example proves that $\frac{5}{4} \leq \rho$. Let's show that $\rho \leq 2.7$. In the prove of a bound for Coffman's algorithm ([5])

$$FF(I) - p_{max} \leq 1.7 \cdot SP(I)$$

(p_{max} stands for greatest length of rectangles) there was used a "areatic" lower bound for optimal value (summary area of rectangle divided by the resource allocation amount) that is also a lower bound for RCPSP optimum. Thus,

$$FF(I) - p_{max} \leq 1.7 \cdot RCPS(I)$$

and

$$SP(I) \leq FF(I) \leq p_{max} + 1.7 \cdot RCPS(I) \leq 2.7 \cdot RCPS(I) \forall I$$

that gives $\rho \leq 2.7$.

Finding of exact value for ρ is interesting since it will show how efficient will be the usage of strip packing algorithms for RCPSP.

Such analysis can be performed for particular algorithms independently. For example, it can be shown that asymptotically optimal algorithm for SP problem with random input (on some class of distributions of input data) is also asymptotically optimal for the corresponding case of RCPSP. Moreover, it can be extended on the case of special "multi-project" kind of precedence graph.

References

1. Gimadi E, Sevastianov S, Zalyubovsky S (2000) Polynomial solvability of project scheduling problems with storable resources and due dates. *Discrete analysis and operations research*, 7, 2: 9–34
2. Gimadi E, Sevastianov S, Zalyubovsky V (2001) On the Project Scheduling Problem under Stored Resource Constraints. *Proceedings of 8th IEEE International Conference on Emerging Technologies and Factory Automation*. Antibes - Juan les Pins, France: 703–706
3. Gimadi E, Zalyubovskii V, Sharygin P (1997) The problem of strip packing: an asymptotically exact approach. *Russian Mathematics (Iz. VUZ)*, Allerton Press Inc., 41, 12: 32–42.
4. Kolish R (1996) Serial and Parallel resource-constrained project scheduling methods revisited: Theory and computation. *European Journal of Operational Research*, 90: 320–333
5. Coffman E, Garey M, Johnson D, Tarjan K (1980) Performance bounds for level-oriented two-dimensional packing algorithms. *SIAM Journal on Computing*, 9, 4: 804–826
6. Baker B, Brown D, Kartseff H (1981) A $5/4$ algorithm for two-dimensional packing. *Journal of algorithms*

Simulation and Applied Probability

Risk-Sensitive Optimality Criteria in Markov Decision Processes

Karel Sladký

Institute of Information Theory and Automation, Academy of Sciences of the
Czech Republic, Praha, Czech Republic
sladky@utia.cas.cz

1 Introduction and Notation

The usual optimization criteria for Markov decision processes (e.g. total discounted reward or mean reward) can be quite insufficient to fully capture the various aspects for a decision maker. It may be preferable to select more sophisticated criteria that also reflect variability-risk features of the problem. To this end we focus attention on risk-sensitive optimality criteria (i.e. the case when expectation of the stream of rewards generated by the Markov processes evaluated by an exponential utility function is considered) and their connections with mean-variance optimality (i.e. the case when a suitable combination of the expected total reward and its variance, usually considered per transition, is selected as a reasonable optimality criterion). The research of risk-sensitive optimality criteria in Markov decision processes was initiated in the seminal paper by Howard and Matheson [6] and followed by many other researchers (see e.g. [1, 2, 3, 5, 4, 8, 9, 14]). In this note we consider a Markov decision chain $X = \{X_n, n = 0, 1, \dots\}$ with finite state space $\mathcal{I} = \{1, 2, \dots, N\}$ and a finite set $\mathcal{A}_i = \{1, 2, \dots, K_i\}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Supposing that in state $i \in \mathcal{I}$ action $k \in \mathcal{A}_i$ is selected, then state j is reached in the next transition with a given probability p_{ij}^k and one-stage transition reward r_{ij} will be accrued to such transition.

Suppose that the stream of transition rewards is evaluated by an exponential utility function, say $u^\gamma(\cdot)$, i.e. a utility function with constant risk sensitivity $\gamma \in \mathbb{R}$. Then the utility assigned to the (random) reward ξ is given by

$$u^\gamma(\xi) := \begin{cases} (\text{sign } \gamma) \exp(\gamma\xi), & \text{if } \gamma \neq 0 \\ \xi & \text{for } \gamma = 0. \end{cases} \quad (1)$$

Obviously $u^\gamma(\cdot)$ is continuous and strictly increasing, and convex (resp. concave) for $\gamma > 0$ (resp. $\gamma < 0$). If ξ is a (bounded) random variable then for the corresponding certainty equivalent of the (random) variable ξ , say $Z(\xi)$,

in virtue of the condition $u^\gamma(Z(\xi)) = \mathbf{E}[(\text{sign } \gamma) \exp(\gamma\xi)]$ ($\mathbf{E}$ is reserved for expectation), we immediately get

$$Z(\xi) = \begin{cases} \frac{1}{\gamma} \ln\{\mathbf{E}[\exp(\gamma\xi)]\}, & \text{if } \gamma \neq 0 \\ \mathbf{E}[\xi] & \text{for } \gamma = 0. \end{cases} \quad (2)$$

Observe that if ξ is constant then $Z(\xi) = \xi$, if ξ is nonconstant then by Jensen's inequality

$$\begin{aligned} Z(\xi) &> \mathbf{E}\xi && (\text{if } \gamma > 0 \text{ and the decision maker is risk averse}) \\ Z(\xi) &< \mathbf{E}\xi && (\text{if } \gamma < 0 \text{ and the decision maker is risk seeking}) \\ Z(\xi) &= \mathbf{E}\xi && (\text{if } \gamma = 0 \text{ and the decision maker is risk neutral}) \end{aligned}$$

A (Markovian) policy, say π , controlling the chain is a rule how to select actions in each state. We write $\pi = (f^0, f^1, \dots)$ where $f^n \in \mathcal{A} \equiv \mathcal{A}_1 \times \dots \times \mathcal{A}_N$ for every $n = 0, 1, 2, \dots$ and $f_i^n \in \mathcal{A}_i$ is the decision at the n th transition when the chain X is in state i . A policy which takes at all times the same decision rule, i.e. selects actions only with respect to the current state and hence is fully identified by some decision vector f whose i th element $f_i \in \mathcal{A}_i$, is called stationary. Stationary policy $\pi \sim (f)$ then completely identifies the transition probability matrix $\mathbf{P}(f)$ along with the one-stage expected reward vector $\mathbf{r}(f)$. Observe that the i th row of $\mathbf{P}(f)$, denoted $\mathbf{p}_i(f)$, has elements $p_{i1}^{f_i}, \dots, p_{iN}^{f_i}$ and that $\mathbf{P}^*(f) = \lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} [\mathbf{P}(f)]^k$ exists. Similarly the i th element of $\mathbf{r}(f)$ denotes the one-stage expected value $r_i^{f_i} = \sum_{k=0}^{n-1} p_{ij}^{f_i} r_{ij}$.

Let $\xi_n = \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}$ be the stream of transition rewards received in the n next transitions of the considered Markov chain X , and similarly let $\xi^{(m,n)}$ be reserved for the total (random) reward obtained from the m th up to the n th transition (obviously, $\xi_n = r_{X_0, X_1} + \xi^{(1,n)}$). In case that $\gamma \neq 0$ then $u^\gamma(\xi_n) := (\text{sign } \gamma) e^{\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}}$ is the (random) utility assigned to ξ_n , and $Z(\xi_n) = \frac{1}{\gamma} \ln \{\mathbf{E}[e^{\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}}] \}$ is its certainty equivalent. Obviously, if $\gamma = 0$ then $u^\gamma(\xi_n) = \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}$ and $Z(\xi_n) = \mathbf{E}[\sum_{k=0}^{n-1} r_{X_k, X_{k+1}}]$.

Supposing that the chain starts in state $X_0 = i$ and policy $\pi = (f^n)$ is followed, then if $\gamma \neq 0$ for the expected utility in the n next transitions and the corresponding certainty equivalent we have ($\mathbf{E}_i^\pi$ denotes expectation if policy π is followed and $X_0 = i$)

$$U_i^\pi(\gamma, 0, n) := \mathbf{E}_i^\pi[u^\gamma(\xi_n)] = (\text{sign } \gamma) \mathbf{E}_i^\pi[\exp(\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}})] \quad (3)$$

$$Z_i^\pi(\gamma, 0, n) := \frac{1}{\gamma} \ln \{ \mathbf{E}_i^\pi[\exp(\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}})] \}. \quad (4)$$

In what follows we shall often abbreviate $U_i^\pi(\gamma, 0, n)$ (resp. $Z_i^\pi(\gamma, 0, n)$) by $U_i^\pi(\gamma, n)$ (resp. $Z_i^\pi(\gamma, n)$). Similarly $\mathbf{U}^\pi(\gamma, n)$ (resp. $\mathbf{Z}^\pi(\gamma, n)$) is reserved for

the vector of expected utilities (resp. certainty equivalents) whose i th element equals $U_i^\pi(\gamma, n)$ (resp. $Z_i^\pi(\gamma, n)$).

In this note we focus attention on the asymptotic behavior of the expected utility and the corresponding certainty equivalents. In particular, we shall study their properties if the risk sensitivity $\gamma > 0$ is close to zero. This will enable to compare this type of risk sensitivity criteria with the mean variance optimality.

2 Asymptotic Behaviour of Expected Utilities

Let $q_{ij}^{f_i} := p_{ij}^{f_i} e^{\gamma r_{ij}}$. Conditioning in (3) on X_1 (observe that $u^\gamma(\xi_n) = \mathbb{E}[u^\gamma(r_{X_0, X_1}) \cdot u^\gamma(\xi^{(1, n)}) | X_1 = j]$) from (3) we immediately get

$$U_i^\pi(\gamma, 0, n) = \sum_{j \in \mathcal{I}} q_{ij}^{f_i^0} U_j^\pi(\gamma, 1, n) \quad \text{with} \quad U_i^\pi(\gamma, n, n) = 1 \quad (5)$$

or in vector notation

$$\mathbf{U}^\pi(\gamma, 0, n) = \mathbf{Q}(f^0) \mathbf{U}^\pi(\gamma, 1, n) \quad \text{with} \quad \mathbf{U}^\pi(\gamma, n, n) = \mathbf{e} \quad (6)$$

where the ij th entry of the $N \times N$ matrix $\mathbf{Q}(f)$ equals $q_{ij}^{f_i} = p_{ij}^{f_i} e^{\gamma r_{ij}}$ and $\mathbf{e}$ is a unit (column) vector.

Iterating (6) we get if policy $\pi = (f^n)$ is followed

$$\mathbf{U}^\pi(\gamma, n) = \mathbf{Q}(f^0) \cdot \mathbf{Q}(f^1) \cdot \dots \cdot \mathbf{Q}(f^{n-1}) \mathbf{e}. \quad (7)$$

Observe that $\mathbf{Q}(f)$ is a nonnegative matrix, and by the Perron–Frobenius theorem the spectral radius $\rho(f)$ of $\mathbf{Q}(f)$ is equal to the maximum positive eigenvalue of $\mathbf{Q}(f)$. Moreover, if $\mathbf{Q}(f)$ is irreducible (i.e. if $\mathbf{P}(f)$ is irreducible) the corresponding (right) eigenvector $\mathbf{v}(f)$ can be selected strictly positive, i.e.

$$\rho(f) \mathbf{v}(f) = \mathbf{Q}(f) \mathbf{v}(f) \quad \text{with} \quad \mathbf{v}(f) > 0. \quad (8)$$

Moreover, under the above irreducibility condition it can be shown (cf. e.g. [6], [9]) that there exists decision vector $f^* \in \mathcal{A}$ such that

$$\mathbf{Q}(f) \mathbf{v}(f^*) \leq \rho(f^*) \mathbf{v}(f^*) = \mathbf{Q}(f^*) \mathbf{v}(f^*), \quad (9)$$

$$\rho(f) \leq \rho(f^*) \equiv \rho^* \quad \text{for all} \quad f \in \mathcal{A}. \quad (10)$$

In words, $\rho(f^*) \equiv \rho^*$ is the maximum possible eigenvalue of $\mathbf{Q}(f)$ over all $f \in \mathcal{A}$.

Furthermore, (9), (10) still hold even for reducible matrices if for suitable labelling of states (i.e. for suitably permuting rows and corresponding columns regardless of the selected $f \in \mathcal{A}$) it is possible to decompose $\mathbf{Q}(f)$ such that

$$\mathbf{Q}(f) = \begin{bmatrix} \mathbf{Q}_{(00)}(f) & \mathbf{Q}_{(01)}(f) & \dots & \mathbf{Q}_{(0r)}(f) \\ \mathbf{0} & \mathbf{Q}_{(11)}(f) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q}_{(rr)}(f) \end{bmatrix} \quad (11)$$

where the spectral radius $\rho_i(f)$ of every irreducible class of $\mathbf{Q}_{(ii)}(f)$ (where $i = 1, \dots, r$) is nongreater than $\rho(f^*)$ (with $\rho_i(f^*) = \rho(f^*)$) and the spectral radius of (possibly reducible) $\mathbf{Q}_{(00)}(f)$ is less than $\rho(f^*)$ and at least some $\mathbf{Q}_{(0j)}(f)$ is nonvanishing.

In case that $\mathbf{Q}(f)$'s are arbitrary reducible nonnegative matrices, it is also possible to extend (9), (10) to this more general case (see Rothblum and Whittle [8], Sladký [10], Whittle [13], and Zijm [14]). In general (after suitable permutations of rows and corresponding columns) every $\mathbf{Q}(f)$ be written in a fixed block-triangular form (independent of $f \in \mathcal{A}$) whose diagonal blocks $\mathbf{Q}_{ii}(f)$ ($i = 1, \dots, s$) with spectral radii $\sigma_i(f)$ (with $\sigma_1(f) \geq \sigma_2(f) \geq \dots \geq \sigma_s(f)$) fulfilling conditions (9), (10), (11) are the largest matrices within the class $\mathbf{Q}(f)$ with $f \in \mathcal{A}$ having positive right eigenvectors corresponding to their spectral radii.

Iterating (9) and using (10) we can immediately conclude that for any policy $\pi = (f^n)$

$$\prod_{k=0}^n \mathbf{Q}(f^k) \mathbf{v}(f^*) \leq (\mathbf{Q}(f^*))^n \mathbf{v}(f^*) = (\rho^*)^n \mathbf{v}(f^*) \quad (12)$$

and hence the asymptotic behaviour of $\mathbf{U}^\pi(\gamma, n)$ (or of $\mathbf{U}^\pi(\gamma, m, n)$ if m is fixed) heavily depends of on $\rho(f^*)$.

Suppose that γ is selected such that $\rho(f^*) < 1$ for all $f \in \mathcal{A}$. Then by (12) $\lim_{n \rightarrow \infty} \prod_{k=0}^n \mathbf{Q}(f^k) = \mathbf{0}$ and by (7), (12) $\mathbf{U}_i^\pi(\gamma, n) \rightarrow 0$ (or $\mathbf{U}^\pi(\gamma, n) \rightarrow \mathbf{0}$) as $n \rightarrow \infty$ and the convergence is geometrical.

For $\gamma = 0$ we have $\mathbf{Q}(f) = \mathbf{P}(f)$, the spectral radius of $\mathbf{Q}(f)$ equals one, and the corresponding right eigenvector $\mathbf{v}(f^*)$ is a constant vector. Then $\mathbf{U}^\pi(\gamma, n) \rightarrow \mathbf{P}^*(f) \mathbf{e}$, a constant vector. Observe that in (11) $\mathbf{Q}_{(ii)}(f)$ with $i \geq 1$ correspond to recurrent classes of $\mathbf{P}(f)$ and $\mathbf{Q}_{(00)}(f)$ corresponds to transient states of $\mathbf{P}(f)$. If $r = 1$ we have the so-called unichain model.

Moreover, let for the case with $\gamma = 0$ the vectors $\mathbf{R}^\pi(n)$, $\mathbf{S}^\pi(n)$, and $\mathbf{V}^\pi(n)$ denote the first moment, the second moment and the variance of the (random) total reward $\xi_n = \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}$ obtained in the n first transitions of the considered Markov chain X if policy $\pi = (f^n)$ is followed, given the initial state $X_0 = i$. More precisely, for the elements of $\mathbf{R}^\pi(n)$, $\mathbf{S}^\pi(n)$, and $\mathbf{V}^\pi(n)$ we have

$$R_i^\pi(n) = \mathbb{E}_i^\pi[\xi_n], \quad S_i^\pi(n) = \mathbb{E}_i^\pi[\xi_n^2], \quad V_i^\pi(n) = \sigma_i^{2, \pi}[\xi_n]$$

where $\mathbb{E}_i^\pi$, $\sigma_i^{2, \pi}$ are standard symbols for expectation and variance if policy π is selected and $X_0 = i$. Since for any integers $m < n$ $\xi_n = \xi_m + \xi^{(m, n)}$ and

$$[\xi_n]^2 = [\xi_m]^2 + 2 \cdot \xi_m \cdot \xi^{(m,n)} + [\xi^{(m,n)}]^2 \text{ we get}$$

$$\mathbb{E}_i^\pi \xi_n = \mathbb{E}_i^\pi \xi_m + \mathbb{E}_i^\pi \xi^{(m,n)} \quad (13)$$

$$\mathbb{E}_i^\pi [\xi_n]^2 = \mathbb{E}_i^\pi [\xi_m]^2 + 2 \cdot \mathbb{E}_i^\pi [\xi_m \cdot \xi^{(m,n)}] + \mathbb{E}_i^\pi [\xi^{(m,n)}]^2. \quad (14)$$

In particular, for $m = 1$, $n := n + 1$ if stationary policy $\pi \sim (f)$ is followed we have

$$R_i^\pi(n+1) = r_i^{f_i} + \sum_{j \in \mathcal{I}} p_{ij}^{f_i} \cdot R_j^\pi(n) \quad (15)$$

$$S_i^\pi(n+1) = \sum_{j \in \mathcal{I}} p_{ij}^{f_i} \cdot \{[r_{ij}]^2 + 2 \cdot r_{ij} \cdot R_j^\pi(n)\} + \sum_{j \in \mathcal{I}} p_{ij}^{f_i} \cdot S_j^\pi(n). \quad (16)$$

Since $V_i^\pi(\cdot) = S_i^\pi(\cdot) - [R_i^\pi(\cdot)]^2$ by (15), (16) we arrive after some algebra at

$$V_i^\pi(n+1) = \sum_{j \in \mathcal{I}} p_{ij}^{f_i} \cdot \{[r_{ij} + R_j^\pi(n)]^2\} - [R_i^\pi(n+1)]^2 + \sum_{j \in \mathcal{I}} p_{ij}^{f_i} \cdot V_j^\pi(n). \quad (17)$$

If we restrict attention on unichain models it is well known (see e.g. [7]) that there exist vector $\mathbf{w}^\pi$, constant vector $\mathbf{g}^\pi$ and vector $\boldsymbol{\varepsilon}(n)$ (where all elements of $\boldsymbol{\varepsilon}(n)$ converge to zero geometrically) such that

$$\mathbf{R}^\pi(n) = \mathbf{g}^\pi \cdot n + \mathbf{w}^\pi + \boldsymbol{\varepsilon}(n) \Rightarrow \lim_{n \rightarrow \infty} n^{-1} \mathbf{R}^\pi(n) = \mathbf{g}^\pi = \mathbf{P}^*(f) \cdot \mathbf{r}(f). \quad (18)$$

The constant vector $\mathbf{g}^\pi$ along with vectors $\mathbf{w}^\pi$ are uniquely determined by

$$\mathbf{w}^\pi + \mathbf{g}^\pi = \mathbf{r}(f) + \mathbf{P}(f) \cdot \mathbf{w}^\pi, \quad \mathbf{P}^*(f) \cdot \mathbf{w}^\pi = \mathbf{0}. \quad (19)$$

Using these facts it can be shown (see [12]) that the growth rate of $V_i^\pi(n)$ is asymptotically linear in n .

Finally, let γ be selected such that the spectral radius ρ^* of $\mathbf{Q}(f^*)$ is greater than unity. Then $U_i^\pi(\gamma, n) \rightarrow \infty$, however the growth rate of $U_i^\pi(\gamma, n)$ (i.e. is the ratio $U_i^\pi(\gamma, n+1)$ to $U_i^\pi(\gamma, n)$) for $n \rightarrow \infty$ is nongreater than ρ^* . Introducing $\tilde{\mathbf{Q}}(f) = (\rho^*)^{-1} \mathbf{Q}(f)$, for the vector of normalized utilities $\tilde{U}^\pi(\gamma, n) := (\rho^*)^{-n} \cdot U^\pi(\gamma, n)$ (with elements $\tilde{U}_i^\pi(\gamma, n) = (\rho^*)^{-n} U_i^\pi(\gamma, n)$) we have

$$\tilde{U}^\pi(\gamma, n+1) = \tilde{\mathbf{Q}}(f) \tilde{U}^\pi(\gamma, n) \quad \text{with} \quad \tilde{U}^\pi(\gamma, 0) = \mathbf{e} \quad (20)$$

and its asymptotic behaviour is similar to the case with $\gamma = 0$ (cf. [11]).

3 Small Risk Aversion Coefficient

Now we focus our attention on the case where the risk aversion coefficient $\gamma > 0$ tends to zero and similarly ρ^* tends to unity. To this end, consider for $\gamma > 0$ Taylor's expansion of (3) around $\gamma = 0$. We immediately get for stationary policy $\pi \sim (f)$

$$\begin{aligned}
U_i^\pi(\gamma, n) = & (\text{sign } \gamma) \{ 1 + \gamma \cdot \mathbf{E}_i^\pi \left[\sum_{k=0}^{n-1} r_{X_k, X_{k+1}} \right] \\
& + \frac{1}{2} (\gamma)^2 \cdot \mathbf{E}_i^\pi \left[\sum_{k=0}^{n-1} r_{X_k, X_{k+1}} \right]^2 + o(\gamma^2) \}. \quad (21)
\end{aligned}$$

In what follows we set $\rho^* = \rho^*(\gamma)$ to stress the dependence of ρ^* on γ and consider for $\gamma > 0$ Taylor's expansion of $\rho^*(\gamma)$ around $\gamma = 0$. Then we get

$$\rho^*(\gamma) = 1 + \rho^{(1)}(f) \gamma + \frac{1}{2} \rho^{(2)}(f) \gamma^2 + o(\gamma^2), \quad (22)$$

where $\rho^{(1)}(f)$, resp. $\rho^{(2)}(f)$ is the spectral radius of the (nonnegative) matrix $\mathbf{Q}^{(1)}(f)$, resp. $\mathbf{Q}^{(2)}(f)$, with elements $q_{ij}^{(1)} = p_{ij}^{f_i} \cdot r_{ij}$, resp. $q_{ij}^{(2)} = p_{ij}^{f_i} \cdot [r_{ij}]^2$.

Hence, on comparing (22) with (17) we can conclude that for sufficiently small $\gamma > 0$ optimality conditions for risk-sensitivity and mean-variance optimality are the same.

Acknowledgement. The research was supported by the Grant Agency of the Czech Republic under Grants 402/05/0115 and 402/04/1294.

References

1. Bielecki TD, Hernández-Hernández TD, Pliska SR (1999) Risk-sensitive control of finite state Markov chains in discrete time, with application to portfolio management. *Math Methods Oper Res* 50:167–188
2. Cavazos-Cadena R, Montes-de-Oca R (2003) The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space. *Math Oper Res* 28:752–756
3. Cavazos-Cadena R (2003) Solution to the risk-sensitive average cost optimality equation in a class of Markov decision processes with finite state space. *Math Methods Oper Res* 57:253–285
4. Jaquette SA (1976) A utility criterion for Markov decision processes. *Manag Sci* 23:43–49
5. Hinderer K, Waldmann KH (2003) The critical discount factor for finite Markovian decision processes with an absorbing set. *Math Methods Oper Res* 57:1–19
6. Howard RA, Matheson J (1972) Risk-sensitive Markov decision processes. *Manag Sci* 23:356–369
7. Puterman ML (1994) Markov decision processes – discrete stochastic dynamic programming. Wiley, New York
8. Rothblum UG, Whittle P (1982) Growth optimality for branching Markov decision chains. *Math Oper Res* 7:582–601
9. Sladký K (1976) On dynamic programming recursions for multiplicative Markov decision chains. *Math Programming Study* 6:216–226
10. Sladký K (1980) Bounds on discrete dynamic programming recursions I. *Kybernetika* 16: 526–547

11. Sladký K (1981) On the existence of stationary optimal policies in discrete dynamic programing. *Kybernetika* 17:489–513
12. Sladký K (2005) On mean reward variance in semi-Markov processes. *Math Methods Oper Res* 62:387–397
13. Whittle P (1983) Optimization over time – dynamic programming and stochastic control. Volume II, Chapter 35, Wiley, Chichester
14. Zijm WHM (1983) Nonnegative matrices in dynamic programming. *Mathematical Centre Tract*, Amsterdam

Trading Regions Under Proportional Transaction Costs

Karl Kunisch¹ and Jörn Sass²

¹ Institute of Mathematics, Karl-Franzens University Graz, Austria
`karl.kunisch@uni-graz.at`

² RICAM, Austrian Academy of Sciences, Linz, Austria
`joern.sass@oeaw.ac.at`

Summary. In the Black-Scholes model optimal trading for maximizing expected power utility under proportional transaction costs can be described by three intervals B , NT , S : If the proportion of wealth invested in the stocks lies in B , NT , S , then buying, not trading and selling, respectively, are optimal. For a finite time horizon, the boundaries of these trading regions depend on time and on the terminal condition (liquidation or not). Following a stochastic control approach, one can derive parabolic variational inequalities whose solution is the value function of the problem. The boundaries of the active sets for the different inequalities then provide the boundaries of the trading regions. We use a duality based semi-smooth Newton method to derive an efficient algorithm to find the boundaries numerically.

1 Trading Without Transaction Costs

The continuous-time Black Scholes model consists of one bond or bank account and one stock with prices $(P_0(t))_{t \in [0, T]}$ and $(P_1(t))_{t \in [0, T]}$ which for interest rate $r \geq 0$, trend $\mu \in \mathbb{R}$, and volatility $\sigma > 0$ evolve according to

$$dP_0(t) = P_0(t) r dt, \quad dP_1(t) = P_1(t) (\mu dt + \sigma dW(t)), \quad P_0(0) = P_1(0) = 1,$$

where $W = (W(t))_{t \in [0, T]}$ is a Brownian motion on a probability space $(\Omega, \mathcal{A}, P)$. Let $\mathcal{F} = (\mathcal{F}_t)_{t \in [0, T]}$ denote the augmented filtration generated by W .

Without transaction costs the trading of an investor may be described by initial capital $\underline{x} > 0$ and risky fraction process $(\eta(t))_{t \in [0, T]}$, where $\eta(t)$ is the fraction of the portfolio value (wealth) which is held in the stocks at time t . The corresponding *wealth process* $(X(t))_{t \in [0, T]}$ is defined self-financing by

$$dX(t) = (1 - \eta(t))X(t)r dt + \eta(t)X(t)(\mu dt + \sigma dW(t)), \quad X(0) = \underline{x}.$$

The utility of terminal wealth $x > 0$ is given by power utility $\frac{1}{\alpha} x^\alpha$ for any $\alpha < 1$, $\alpha \neq 0$. The parameter α models the preferences of an investor. The limiting case $\alpha \rightarrow 0$ corresponds to logarithmic utility i.e. maximizing the expected rate of return,

$\alpha > 0$ corresponds to less risk averse and $\alpha < 0$ to more risk averse utility functions. Merton showed that for logarithmic ($\alpha = 0$) and power utility the optimal trading strategy is given by a constant optimal risky fraction

$$\eta(t) = \hat{\eta}, \quad t \in [0, T], \quad \text{where} \quad \hat{\eta} = \frac{1}{1 - \alpha} \frac{\mu - r}{\sigma^2}. \quad (1)$$

2 Proportional Transaction Costs

To keep the risky fraction constant like in (1) involves continuous trading which, under transaction costs, is no longer adequate. For possible cost structures see e.g. [5, 7, 8]. We consider proportional costs $\gamma \in (0, 1)$ corresponding to the proportion of the traded volume which has to be paid as fees.

For suitable infinite horizon criteria solution of the corresponding Hamilton-Jacobi-Bellmann equation (HJB) leads to a characterization of the optimal wealth process as a diffusion reflected at the boundaries of a cone, see [3, 9]. When reaching the boundaries of the cone, infinitesimal trading occurs in such a way that the wealth process just stays in the cone. The cone corresponds to an interval for the risky fraction process. The existence of a viscosity solution for the HJB equation for finite time horizon is shown in [1] and numerically treated in [10] using a finite difference method.

Now let us fix costs $\gamma \in (0, 1)$ and parameters $\alpha < 1$, $\alpha \neq 0$, r , μ , σ such that $\hat{\eta} \in (0, 1)$. The trading policy can be described by two increasing processes $(L(t))_{t \in [0, T]}$ and $(M(t))_{t \in [0, T]}$ representing the cumulative purchases and sales of the stock. We require that these are right-continuous, $\mathcal{F}$ -adapted, and start with $L(0-) = M(0-) = 0$. Transaction fees are paid from the bank account. Thus the dynamics of the controlled wealth processes $(X_1(t))_{t \in [0, T]}$ and $(X_0(t))_{t \in [0, T]}$, corresponding to the amount of money on the bank account and the amount invested in the stocks, are

$$\begin{aligned} dX_0(t) &= rX_0(t) dt - (1 + \gamma) dL(t) + (1 - \gamma) dM(t), \\ dX_1(t) &= \mu X_1(t) dt + \sigma X_1(t) dW(t) + dL(t) - dM(t). \end{aligned}$$

The objective is the maximization of expected utility at the terminal trading time T , now over all control processes $(L(t))_{t \in [0, T]}$ and $(M(t))_{t \in [0, T]}$ which satisfy the conditions above and for which the wealth processes X_0 and X_1 stay positive and the total wealth strictly positive i.e. $(X_0(t), X_1(t)) \in \mathcal{D} := \mathbb{R}_+^2 \setminus \{(0, 0)\}$, $t \in [0, T]$. So suppose $(\underline{x}_0, \underline{x}_1) = (X_0(0-), X_1(0-)) \in \mathcal{D}$. We distinguish the maximization of expected utility for the terminal total wealth,

$$\tilde{J}(t, x_0, x_1) = \sup_{(L, M)} \mathbb{E} \left[\frac{1}{\alpha} (X_0(T) + X_1(T))^\alpha \mid X_0(t) = x_0, X_1(t) = x_1 \right],$$

and of the terminal wealth after liquidating the position in the stocks,

$$J(t, x_0, x_1) = \sup_{(L, M)} \mathbb{E} \left[\frac{1}{\alpha} (X_0(T) + (1 - \gamma)X_1(T))^\alpha \mid X_0(t) = x_0, X_1(t) = x_1 \right].$$

We always assume $\tilde{J}(0, x_0, x_1) < \infty$ for all $(x_0, x_1) \in \mathcal{D}$.

Theorem 1. *J is concave, continuous, and a viscosity solution of*

$$\max\{J_t + \mathcal{A}J, -(1 + \gamma)J_{x_0} + J_{x_1}, (1 - \gamma)J_{x_0} - J_{x_1}\} = 0 \quad (2)$$

on $[0, T] \times \mathcal{D}$ with $J(T, x_0, x_1) = \frac{1}{\alpha}(x_0 + (1 - \gamma)x_1)^\alpha$. $J_t, J_{x_0}, J_{x_1}, J_{x_1, x_1}$ denote the partial derivatives of $J = J(t, x_0, x_1)$ and the differential operator $\mathcal{A}$ (generator of (X_0, X_1)) is defined by

$$\mathcal{A}h(x_0, x_1) = r x_0 h_{x_0}(x_0, x_1) + \mu x_1 h_{x_1}(x_0, x_1) + \frac{1}{2} \sigma^2 x_1^2 h_{x_1, x_1}(x_0, x_1)$$

for all smooth functions h . Further J is unique in the class of continuous functions satisfying $|h(t, x_0, x_1)| \leq K(1 + (x_0^2 + x_1^2)^\alpha)$ for all $(x_0, x_1) \in \mathcal{D}$, $t \in [0, T]$, and some constant K .

The proof in [1] is based on the derivation of a weak dynamic programming principle leading to the HJB (2). The uniqueness is shown following the Ishii technique, see [2]. Using $\tilde{J}(T, x_0, x_1) = \frac{1}{\alpha}$, the argument is the same for $\tilde{J}$.

The variational inequalities in (2) are active, if it is optimal not to trade, to buy stocks, and to sell stocks, respectively. At time t , $\mathcal{D}$ can be split into the buy region $B'(t)$, the sell region $S'(t)$, and the no trading region $NT'(t)$,

$$\begin{aligned} B'(t) &= \{(x_0, x_1) \in \mathcal{D} : -(1 + \gamma)J_{x_0}(t, x_0, x_1) + J_{x_1}(t, x_0, x_1) = 0\}, \\ S'(t) &= \{(x_0, x_1) \in \mathcal{D} : (1 - \gamma)J_{x_0}(t, x_0, x_1) - J_{x_1}(t, x_0, x_1) = 0\}, \\ NT'(t) &= \mathcal{D} \setminus (B'(t) \cup S'(t)). \end{aligned}$$

If $x_0 = 0$ ($x_1 = 0$) we should exclude the second (third) inequality in (2) since buying (selling) is not admissible. But due to the $\hat{\eta} \in (0, 1)$ we expect that (x_0, x_1) lies for all t in $NT'(t) \cup S'(t)$ if $x_0 = 0$ and in $NT'(t) \cup B'(t)$ if $x_1 = 0$, cf. [9]. Thus we did not specify the different cases in Theorem 1.

3 Reduction to a One-Dimensional Problem

We use a transformation to the risky fractions, different to e.g. [10] where the fractions modified by the transaction costs are used. From the definition of J and $\tilde{J}$ one verifies directly that on $[0, T] \times \mathcal{D}$

$$J(t, x_0, x_1) = x^\alpha J(t, \frac{x_0}{x}, \frac{x_1}{x}), \quad \tilde{J}(t, x_0, x_1) = x^\alpha \tilde{J}(t, \frac{x_0}{x}, \frac{x_1}{x}), \quad x = x_0 + x_1.$$

So it is enough to look at the risky fractions $y = \frac{x_1}{x}$. Introducing

$$V(t, y) = J(t, 1 - y, y), \quad \tilde{V}(t, y) = \tilde{J}(t, 1 - y, y), \quad y \in [0, 1]$$

we get for $J(t, x_0, x_1) = (x_0 + x_1)^\alpha V(t, \frac{x_1}{x_0 + x_1})$ (writing $x = x_0 + x_1$)

$$\begin{aligned} J_t &= x^\alpha V_t, \quad J_{x_0} = x^{\alpha-1}(\alpha V - y V_y), \quad J_{x_1} = x^{\alpha-1}(\alpha V + (1 - y) V_y), \\ J_{x_1, x_1} &= x^{\alpha-2}((1 - y)^2 V_{y, y} - 2(1 - y)(1 - \alpha) V_y - \alpha(1 - \alpha) V). \end{aligned}$$

Plugging this into (2) we have to solve

$$\max\{V_t + \mathcal{L}V, \mathcal{L}_B V, \mathcal{L}_S V\} = 0 \quad (3)$$

with $V(T, y) = \frac{1}{\alpha}(1 - \gamma y)^\alpha$, $y \in [0, 1]$, and operators

$$\begin{aligned}\mathcal{L}h(y) &= \alpha \left(r + (\mu - r)y - \frac{1}{2}(1 - \alpha)\sigma^2 y^2 \right) \\ &\quad + \left((\mu - r)(1 - y) - (1 - \alpha)\sigma^2 y(1 - y) \right) y h_y(y) + \frac{1}{2}\sigma^2 y^2 (1 - y)^2 h_{y,y}(y), \\ \mathcal{L}_B h(y) &= (1 + \gamma y)h_y - \alpha \gamma h, \quad \mathcal{L}_S h(y) = -(1 - \gamma y)h_y - \alpha \gamma h.\end{aligned}$$

The same applies to $\tilde{V}$ using the terminal condition $\tilde{V}(T, y) = \frac{1}{\alpha}$ instead. The trading regions are now given by $B(t) = \{y \in [0, 1] : \mathcal{L}_B V(t, y) = 0\}$, $S(t) = \{y \in [0, 1] : \mathcal{L}_S V(t, y) = 0\}$ and $NT(t) = [0, 1] \setminus (B(t) \cup S(t))$ corresponding to buying, selling and not trading, respectively. On $B(t)$ and $S(t)$ we hence know that V satisfies as a solution of $\mathcal{L}_B V = 0$ and $\mathcal{L}_S V = 0$

$$V(t, y) = C_B(t)(1 + \gamma y)^\alpha \quad \text{and} \quad V(t, y) = C_S(t)(1 - \gamma y)^\alpha. \quad (4)$$

As proven in many cases we assume that $B(t)$, $N(t)$, $S(t)$ are intervals. So they can be described by their boundaries

$$a(t) = \inf NT(t), \quad b(t) = \sup NT(t). \quad (5)$$

From the condition $\hat{\eta} \in (0, 1)$ it is reasonable to expect that $NT(t) \neq \emptyset$ for all $t \in [0, T]$. But since borrowing and short selling are not allowed, this might not be true for all $B(t)$ and $S(t)$. If that happens we need boundary conditions different from (4) to solve for V on $NT(t)$. These are

$$V_i(t, 0) + \alpha r V(t, 0) = 0, \quad \text{if } 0 \in NT(t), \quad (6)$$

$$V_i(t, 1) + \alpha(\mu - \frac{1}{2}(1 - \alpha)\sigma^2)V(t, 1) = 0, \quad \text{if } 1 \in NT(t). \quad (7)$$

4 A Semi-Smooth Newton Method

The algorithm we present to solve (3) is based on a primal-dual active set strategy, see e.g. [4] where the relationship to semismooth Newton methods is explored, or for its parabolic version cf. [6] where it is also applied to find the exercise boundary for an American option. Here we face two free boundaries and a different type of constraints and hence have to adapt the algorithm. A more detailed analysis including convergence and existence of the Lagrange multipliers is work in progress and deferred to a future publication. However, Example 1 below shows that the algorithm can work efficiently.

Problem (3) is equivalent to solving

$$V_t + \mathcal{L}V + \ell_B + \ell_S = 0, \quad (8)$$

$$\mathcal{L}_B V \leq 0, \ell_B \geq 0, \ell_B \mathcal{L}_B V = 0, \quad \mathcal{L}_S V \leq 0, \ell_S \geq 0, \ell_S \mathcal{L}_S V = 0. \quad (9)$$

The two complementarity problems in (9) can be written as

$$\ell_B = \max\{0, \ell_B + c \mathcal{L}_B V\}, \quad \ell_S = \max\{0, \ell_S + c \mathcal{L}_S V\} \quad (10)$$

for any constant $c > 0$. So we have to solve (8), (10). At T the trading regions are given by $S(T) = [0, 1]$ for V and $NT(T) = [0, 1]$ for $\tilde{V}$. We split $[0, T]$ in N intervals and go backwards in time with $t_N = T$, $t_n = t_{n+1} - \Delta t$, $\Delta t = T/N$. Having computed $V(t_{n+1}, \cdot)$ and the corresponding regions we use the following algorithm to compute $v = V(t_n, \cdot)$ and $NT(t_n)$:

0. Set $\bar{v} = V(t_{n+1}, \cdot)$, $k = 0$, choose an interval NT_0 in $[0, 1]$, constant $c > 0$.
1. Define the boundaries a_k and b_k of NT_k as in (5).
2. On $[a_k, b_k]$ solve the elliptic problem $\frac{1}{\Delta t}(\bar{v} - v) + \mathcal{L}v = 0$ using the boundary conditions $\mathcal{L}_B v = 0$ if $a_k \notin NT_k$, (6) if $a_k \in NT_k$ (implying $a_k = 0$) and $\mathcal{L}_S v = 0$ if $b_k \notin NT_k$, (7) if $b_k \in NT_k$ (implying $b_k = 1$).
3. If $a_k \neq 0$ define v on $[0, a_k]$ by the first equation in (4). If $b_k \neq 1$ define v on $(b_k, 1]$ by the second equation in (4). Choose C_B and C_S such that v is continuous in a_k and b_k . So $v_{k+1} = v$ is continuously differentiable.
4. Set $\ell_B^{k+1} = -\frac{1}{\Delta t}(\bar{v} - v_{k+1}) - \mathcal{L}v_{k+1}$ on $[0, a_k]$ and $\ell_S^{k+1} = -\frac{1}{\Delta t}(\bar{v} - v_{k+1}) - \mathcal{L}v_{k+1}$ on $[b_k, 1]$ and set them to 0 otherwise.
5. Introduce the active sets

$$B_{k+1} = \{y \in [0, 1] : \ell_B^{k+1}(y) + c \mathcal{L}_B v_{k+1}(y) > 0\},$$

$$S_{k+1} = \{y \in [0, 1] : \ell_S^{k+1}(y) + c \mathcal{L}_S v_{k+1}(y) > 0\}$$

and set $NT_{k+1} = [0, 1] \setminus (B_{k+1} \cup S_{k+1})$. Verify that the interval structure holds and define the boundaries a_{k+1} and b_{k+1} by (5).

6. If $a_{k+1} = a_k$ and $b_{k+1} = b_k$ then STOP; otherwise increase k by 1 and continue with step 1.

Example 1. We consider a money market and a stock with parameters $r = 0$, $\mu = 0.096$, $\sigma = 0.4$, and horizon $T = 1$. We use mesh sizes $\Delta t = 0.01$ and $\Delta y = 0.001$, choose $c = 1$, and at t_{N-1} use $NT_0 = (0.1, 0.8)$, and at all other time steps t_n use $NT_0 = NT(t_{n+1})$. For the utility function we consider both $\alpha = 0.1$ and the more risk averse parameter $\alpha = -1$. These yield without transaction costs optimal risky fractions 0.667 and 0.3 (dotted lines in Figs. 1 and 2). We consider proportional costs $\gamma = 0.01$. In Fig. 1 we look at $\alpha = 0.1$, left-hand at V with liquidation at the end, right-hand at $\tilde{V}$. We see that the liquidation costs we have to pay at T imply that we also trade close to the terminal time, while without liquidation this is never optimal. In Fig. 2 we plotted the trading regions for the more risk averse parameter $\alpha = -1$, which leads to less holdings in the stock.

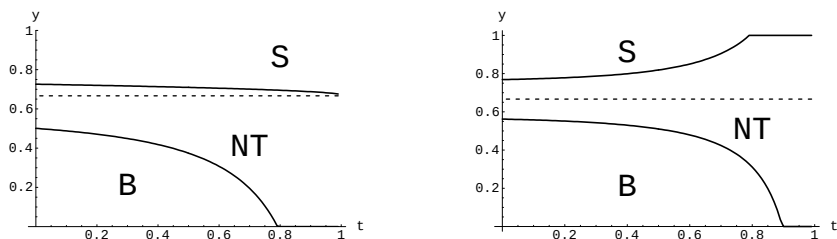


Fig. 1. Trading regions for $\alpha = 0.1$ for V and $\tilde{V}$

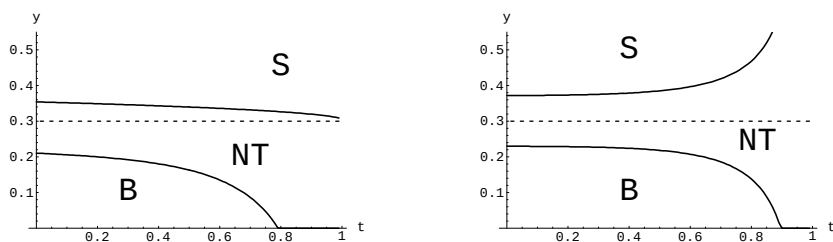


Fig. 2. Trading regions for $\alpha = -1$ for V and $\tilde{V}$

References

1. Akian M, Sulem A, S  quier P (1996) A finite horizon multidimensional portfolio selection problem with singular transactions. *Proceedings of the 34th Conference on Decisions & Control*, New Orleans: 2193–2197
2. Crandall MG, Ishii H, Lions, P-L (1992) User’s guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society* 27: 1–67
3. Davis MHA, Norman AR (1990) Portfolio selection with transaction costs. *Mathematics of Operations Research* 15: 676–713
4. Hinterm  ller M, Ito K, Kunisch K (2003) The primal-dual active set strategy as a semismooth Newton method. *SIAM Journal on Optimization* 13: 865–888
5. Irle A, Sass J (2005) Good portfolio strategies under transaction costs: A renewal theoretic approach. In: do Rosario Grossinho M, Shiryaev AN, Esquivel ML, Oliveira PE (eds) *Stochastic Finance*, Springer, New York: pp. 321–341 .
6. Ito K, Kunisch K (2006) Parabolic variational inequalities: The Lagrange multiplier approach. *Journal de Math  matiques Pures et Appliqu  es* (9) 85: 415–449
7. Korn R (1997) *Optimal Portfolios: Stochastic Models for Optimal Investment in Continuous Time*. World Scientific, Singapore
8. Sass J (2005) Portfolio optimization under transaction costs in the CRR model. *Mathematical Methods of Operations Research* 61: 239–259
9. Shreve SE, Soner HM (1994) Optimal investment and consumption with transaction costs. *Annals of Applied Probability* 4, 609–692
10. Sulem A (1997) Dynamic optimization for a mixed portfolio with transaction costs. In: Rogers LCG, Talay D (eds) *Numerical Methods in Finance*. Cambridge University Press, Cambridge: 165–180

Uniform Random Rational Number Generation

Thomas Morgenstern

FB Automatisierung und Informatik, Hochschule Harz, Wernigerode, Germany
`tmorgenstern@hs-harz.de`

Summary. Classical floating point random numbers fail simple tests when considered as rational numbers.

A uniform random number generator based on a linear congruential generator with output in the rational numbers is implemented within the two computer algebra systems Maple 10 and MuPAD Pro 3.1.

The empirical tests in Knuth's suite of tests show no evidence against the hypothesis that the output samples independent and identically uniformly distributed random variables.

1 Introduction

Randomized algorithms like Monte Carlo simulation require long sequences of random numbers $(x_i)_{i \in \mathbb{N}}$. These sequences are asked to be samples of independent and identically distributed (i.i.d.) random variables $(X_i)_{i \in \mathbb{N}}$. Of special interest are in the open unit interval $(0,1) \subseteq \mathbb{R}$ uniformly distributed random variables $U_i \sim \mathcal{U}(0,1)$.

Random numbers produced by algorithmic random number generators (RNG) are never random, but should appear to be random to the uninitiated. We call these generators pseudorandom number generators [1]. They should pass statistical tests of the hypothesis [2]:

H_0 : the sequence is a sample of i.i.d. random variables with the given distribution.

In fact, no pseudo RNG can pass all statistical tests. So we may say that bad RNGs are those that fail simple tests, whereas good RNGs fail only complicated tests that are very hard hard to find and to run [4, 5].

1.1 Definitions

We follow the definitions in [2, 4, 3]:

Definition 1. A (pseudo-) random number generator (RNG) is a structure $(\mathcal{S}, \mu, t, \mathcal{O}, o)$ where $\mathcal{S}$ is a finite set of states (the state space), μ is a probability distribution on $\mathcal{S}$ used to select the initial state (or seed) s_0 , $t : \mathcal{S} \rightarrow \mathcal{S}$ is the transition function, $\mathcal{O}$ is the output space, and $o : \mathcal{S} \rightarrow \mathcal{O}$ is the output function.

The state of the RNG evolves according to the recurrence $s_i = t(s_{i-1})$, for $i \geq 1$, and the output at step i is $u_i = o(s_i) \in \mathcal{O}$. The output values $u_0, u_1, u_2, \dots$ are called the *random numbers* produced by the RNG.

In order to keep the computation times moderate and to compare our results with the literature we use a generator implemented in Maple 10 [9] and MuPAD Pro 3.1 [10] (and discussed in [4]) and call it LCG10. It is a multiplicative congruential generator ($c = 0$) with multiplier $a = 427\,419\,669\,081$ and modulus $m = 10^{12} - 11 = 999\,999\,999\,989$, i.e. with the recurrence:

$$s_{i+1} := 427\,419\,669\,081 \cdot s_i \bmod 999\,999\,999\,989 \quad . \quad (1)$$

It has period length $\rho = 10^{12} - 12$. To produce values in the unit interval $\mathcal{O} = (0, 1) \subseteq \mathbb{R}$ one usually uses the output function:

$$u_i = o(s_i) := s_i / m \quad .$$

1.2 Problems and Indications

Example 1. Consider a micro wave with frequency 10^{10} Hz. We want to determine the signal energy $E = \int_0^1 \cos(2\pi 10^{10} \times t)^2 dt$ by Monte-Carlo integration. We use the generator Eq. 1 to produce (decimal 10 digits precision floating point) numbers $u_i \in (0, 1)$ and simulate the times $t_i := 2\pi 10^{10} \times u_i$. For n random numbers and $S := \sum_{i=1}^n \cos^2(t_i)$ we expect $E \approx S/n$.

Using $n := 10^6$ and 10 iterations starting with seed 1 we get values in the interval $[0.949, 0.950]$ with mean 0.9501, far from the true result 0.50. Using the same numbers to integrate the cosine $\int_0^1 \cos(2\pi 10^{10} \times t) dt$ we get values in $[0.899, 0.901]$ with mean 0.9000.

These results are due to the fact that mainly natural multipliers of 2π are generated. To see this, we multiply the number u_i by $m = 10^{12}$ or $m = 999\,999\,999\,989$ and treat the result as rational number r_i (e.g. convert it to a rational number using `convert(u, rational, exact)` or `convert(u, rational, 10)` in Maple 10). The smallest common multiplier of the divisors is 1, what is unlikely for true random rational numbers.

2 Random Rational Numbers

We construct a new generator producing rational numbers as output.

2.1 An Uniform Random Rational Number Generator

The new generator RationalLCG10 is defined in Table 1 and is based on LCG10 (Eq. 1). The initial state s_0 is chosen arbitrarily.

Table 1. RationalLCG10

```
s := 427 419 669 081*s mod 999 999 999 989;
q := s;
s := 427 419 669 081*s mod 999 999 999 989;
if (s < q) then p := s else p := q; q := s end if;
return r := p/q;
```

The generator has period length $\rho = 10^{11} - 6$, i.e. half the period length of LCG10. (Note that $p \neq q$, because two consecutive states of a LCG can only be equal if all states are the same.)

2.2 Empirical Tests

We test our generator RationalLCG10 with Knuth's suite of empirical tests (see [1, pp. 61–73]), all implemented in Maple 10 [9] and MuPAD Pro 3.1 [10].

2.2.1 Kolmogorov-Smirnov Equidistribution Test

The Kolmogorov-Smirnov test is described in [1, pp. 48–58]. Starting with seed 1 and $n = 10^5$ numbers, we get in 55 tests the results in Table 2.

Table 2. Kolmogorov-Smirnov test $n = 10^5$

$0\% \leq p^+ < 5\%$	4	$0\% \leq p^- < 5\%$	4
$5\% \leq p^+ < 10\%$	3	$5\% \leq p^- < 10\%$	3
$10\% \leq p^+ \leq 90\%$	42	$10\% \leq p^- \leq 90\%$	41
$90\% < p^+ \leq 95\%$	3	$90\% < p^- \leq 95\%$	4
$95\% < p^+ \leq 100\%$	3	$95\% < p^- \leq 100\%$	3

2.2.2 Chi-Square Equidistribution Test

The Chi-square test is described in [1, pp. 42–47]. Starting with seed 1 and 20 tests we get with $n = 10^5$ random numbers and 20 000 subintervals the occurrences in Table 3(a). With $n = 10^6$ and 200 000 classes we get the results in Table 3(b).

Table 3. Chi-square test: (a) $n = 10^5$; (b) $n = 10^6$

$0\% \leq p < 5\%$	2	$0\% \leq p < 5\%$	1
$5\% \leq p < 10\%$	1	$5\% \leq p < 10\%$	3
$10\% \leq p \leq 90\%$	17	$10\% \leq p \leq 90\%$	15
$90\% < p \leq 95\%$	0	$90\% < p \leq 95\%$	0
$95\% < p \leq 100\%$	0	$95\% < p \leq 100\%$	1

2.2.3 Birthday Spacings Test

The Birthday spacings test is described in [1, pp. 71–72] and discussed in [6, 7, 5, 8]). We consider a sparse case in $t = 2$ dimensions.

For $n = 10^5$ random numbers in 50 000 random vectors we choose $d = 5\,590\,169$ classes per dimension, such that we have $k = 31\,249\,989\,448\,561$ categories in total. For $n = 10^6$ random numbers, we choose $d = 176\,776\,695$ classes per dimension, such that we have $k = 31\,249\,999\,895\,123\,025$ categories. In both cases we expect in average $\lambda = 1$ collisions of birthday spacings.

For LCG10 in 10 tests (and initial state 1) we get more than 3 collisions for the tests in the first case and more than 1 100 collisions in the second case. The probability to get these or more collisions is less than 2% and even less than 10^{-40} in the second case. LCG generators are known to fail this test for $n > \sqrt[3]{p}$ (see [6, 5, 8]). For the generator RationalLCG10 we get the unsuspect results shown in Table 4.

Table 4. Birthday spacing test: (a) $n = 10^5$; (b) $n = 10^6$

$0\% \leq p < 5\%$	2	$0\% \leq p < 5\%$	2
$5\% \leq p < 10\%$	3	$5\% \leq p < 10\%$	2
$10\% \leq p \leq 90\%$	15	$10\% \leq p \leq 90\%$	16
$90\% < p \leq 95\%$	0	$90\% < p \leq 95\%$	0
$95\% < p \leq 100\%$	0	$95\% < p \leq 100\%$	0

2.2.4 Serial Correlation Test

The Serial correlation test computes the auto correlations as described in [1, pp. 72–73]. With $n = 10^6$ random numbers one expects correlations within the range $[-0.002, 0.002]$ in 95% of the tests.

For the generator RationalLCG10 starting from seed 1 with lags from 1 up to 100 in one test we get the results shown in Table 5.

2.3 Improvements

We continue Example 1 using the generator RationalLCG10. In 10 simulations with $n = 10^6$ numbers we get in average 0.500 028 for the Energy and 0.000 386 for the

Table 5. Serial correlation test: $n = 10^6$

$p < -0.003$	0	$0.0 \leq p \leq 0.002$	46
$0.003 \leq p < -0.002$	2	$0.002 < p \leq 0.003$	2
$-0.003 \leq p \leq 0.0$	50	$0.003 < p$	0

Table 6. Monte-Carlo Simulation: (a) $\cos(x)$; (b) $\cos^2(x)$

$-0.0015 \leq \bar{S} \leq -0.0005$	4	$0.4993 \leq \bar{S} < 0.4998$	7
$-0.0005 < \bar{S} \leq +0.0005$	7	$0.4998 < \bar{S} \leq 0.5003$	10
$+0.0005 < \bar{S} \leq +0.0015$	9	$0.5003 < \bar{S} \leq 0.5008$	3

the integral of $\cos$. In Table 6 we find the results of 20 simulations, with average 0.000 270 for the integral of $\cos$ and 0.499 968 for $\cos^2$.

The generator RationalLCG10 produces random numbers with many different denominators. As in Sec. 1.2 we multiply the rational numbers by $m = 999\,999\,999\,989$, take the denominators q_i and their least common multiple d . Plotting $\log_{10}(d)$ against the number of iterations i , we find exponential growth (see Fig. 1).

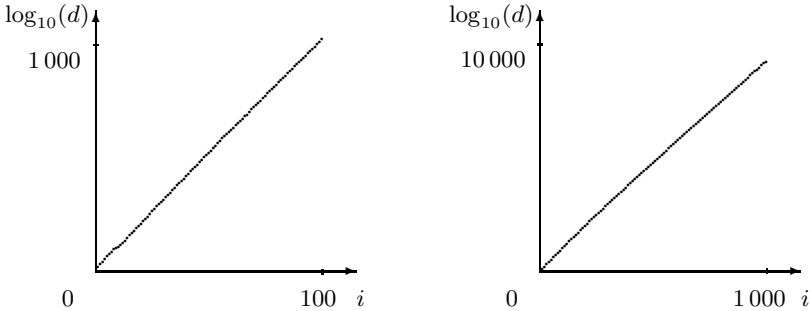


Fig. 1. $\log_{10}$ of the lcm of the denominators of RationalLCG10

3 Conclusions

Using division by the modulus as output function for a linear congruential random number generator can lead to bad results even in simple simulations. This is mainly due to the fact that only a few denominators are generated. A first improvement was suggested in [11].

The new generator RationalLCG10 produces rational random numbers equally distributed in $(0, 1)$. It shows a greater variety of denominators and improved simulation results, compared to the underlying generator LCG10.

The generator RationalLCG10 passes all of Knuth's empirical tests [1] (not all results are presented here due to the lack of space) and the birthday spacings test indicates even further improvements.

References

1. Knuth D E (1998) The art of computer programming. Vol. 2: Seminumerical algorithms. third edition, Addison-Wesley, Reading, Mass.
2. L'Ecuyer P (1994) Uniform random number generation. *Annals of Operations Research* 53:77–120
3. L'Ecuyer P (1998) Random number generation. In: Banks J (ed) *Handbook on Simulation*. John Wiley, Hoboken, NJ.
4. L'Ecuyer P (2004) Random number generation. In: Gentle J E, Härdle W, Mori Y, (eds) *Handbook of computational statistics. Concepts and methods*. Springer, Berlin Heidelberg New York
5. L'Ecuyer P (2001) Software for uniform random number generation: Distinguishing the good and the bad. In: *Proceedings of the 2001 Winter Simulation Conference*. Piscataway NJ., IEEE Press
6. L'Ecuyer P, Hellekalek P (1998) Random number generators: Selection criteria and testing. In: Hellekalek P (ed) *Random and quasi-random point sets*. Springer Lecture Notes in Statistics 138. Springer, New York
7. L'Ecuyer P, Simard R (2001) On the performance of birthday spacings tests with certain families of random number generators. *Mathematics and Computers in Simulation* 55(1-3): 131–137
8. L'Ecuyer P, Simard R, Wegenkittl S (2002) Sparse serial tests of uniformity for random number generators. *SIAM Journal on Scientific Computing* 24(2): 652–668
9. Maple 10 (2005) Maplesoft, a division of Waterloo Maple Inc., www.maplesoft.com
10. MuPAD Pro 3.1 (2005) SciFace Software GmbH& Co.KG, www.sciface.com
11. Morgenstern T (2006) Uniform Random Binary Floating Point Number Generation. In: *Proceedings of the 2. Wernigeröder Automatisierungs- und Informatiktage*. Hochschule Harz, Wernigerode

The Markov-Modulated Risk Model with Investment

Mirko Kötter¹ and Nicole Bäuerle²

¹ Institute for Mathematical Stochastics, University of Hannover, Germany
`koetter@stochastik.uni-hannover.de`

² Institute for Stochastics, University of Karlsruhe, Germany
`baeuerle@stoch.uni-karlsruhe.de`

Summary. We consider Markov-modulated risk reserves which can be invested into a stock index following a geometric Brownian motion. Within a special class of investment policies we identify one which maximizes the adjustment coefficient. A comparison to the compound Poisson case is also given.

1 Introduction and Model

In this paper we combine two features of risk models which have so far only been investigated separately: Markov-modulation (cf. [1]) and maximizing the adjustment coefficient by optimal investment (cf. [6]). In the Markov-modulated Poisson model the premium rate and claim arrivals are determined by a Markov-modulated environment which is described by an irreducible continuous-time Markov process J defined on some finite state space $E = \{1, \dots, d\}$ with intensity matrix $Q = (q_{ij})_{i,j \in E}$. If not stated otherwise, the distribution of J_0 is arbitrary. We only suppose that $\mathbb{P}(J_0 = i) > 0$ for all $i \in E$. J_t can be interpreted as the general economic conditions which are present at time t . J_t influences the premium rate, the arrival intensity of claims and the claim size distribution as follows: The premium income rate at time t is c_{J_t} , i.e. as long as $J_t = i$ we have a linear income stream at rate c_i . Claim arrivals are according to a Poisson process with rate λ_{J_t} . Thus, $N = \{N_t, t \geq 0\}$ is a Markov-modulated Poisson process. A claim U_k which occurs at time t has distribution B_{J_t} , where B_i is some distribution concentrated on $(0, \infty)$. As usual the claim sizes $U_1, U_2, \dots$ are assumed to be conditionally independent given J and μ_i is the finite expectation of B_i .

In our model the insurer has the opportunity to invest into a stock index or say some portfolio whose price process $S := \{S_t, t \geq 0\}$ is modeled by a geometric Brownian motion with dynamics $dS_t = S_t(a dt + b dW_t)$ with $a \in \mathbb{R}$, $b > 0$. K_t denotes the amount of money which the insurer invests into the portfolio at time t and is also allowed to be negative or even larger than the actual wealth for any $t \geq 0$. This fact can be interpreted as the possibility to sell the portfolio short or to borrow an arbitrary amount of money from the bank respectively. The remaining

part of the insurers reserve is invested into a bond which yields no interest. Our aim is to maximize the adjustment coefficient, i.e. the parameter which determines the exponential decay of the ruin probability w.r.t. the initial reserve. In this paper we restrict ourselves to the case where the admissible investment strategies depend on the environment only, i.e. we consider functions $k : E \rightarrow \mathbb{R}$ such that the investment strategy K is given by $K_t = k(J_t)$. This is reasonable since in the paper by [6] the authors show in the model without Markov-modulation that a constant investment strategy maximizes the adjustment coefficient. Thus, we suppose that the wealth process of the insurer is given by

$$Y_t(u, K) = u + \int_0^t c_{J_s} ds - \sum_{k=1}^{N_t} U_k + \int_0^t \frac{K_v}{S_v} dS_v. \quad (1)$$

By applying the time change $\hat{Y}_t := Y_{T(t)}$ with $T(t) := \int_0^t \frac{1}{c_{J_s}} ds$ the ruin probability does not change and we can w.l.o.g. assume that $c(\cdot) \equiv c \in \mathbb{R}$.

In this paper we suppose that all claims have exponential moments, i.e. there exists a possibly infinite constant $r_\infty^{(i)} > 0$ such that the centered moment generating function $h_i(r) := \int_0^\infty e^{rx} dB_i(x) - 1$ is finite for every $r < r_\infty^{(i)}$. It is moreover assumed that $h_i(r) \rightarrow \infty$ as $r \rightarrow r_\infty^{(i)}$. This assumption implies that h_i is increasing, convex and continuous on $[0, r_\infty^{(i)})$ with $h_i(0) = 0$. An important part of this assumption is that $h_i(r) < \infty$ for some $r > 0$. Thus, the tail of the distribution B_i decreases at least exponentially fast.

When applying an investment strategy K , the ruin probability in infinite time is then for $u \geq 0$ defined by

$$\Psi(u, K) = \mathbb{P} \left(\inf_{t \geq 0} Y_t(u, K) < 0 \right).$$

If $\tau(u, K) := \inf\{t > 0; Y_t(u, K) < 0\}$ is the time of ruin, then $\Psi(u, K) = \mathbb{P}(\tau(u, K) < \infty)$. If we denote by $\pi = (\pi)_{i \in E}$ the stationary distribution of J (which exists and is unique since J is irreducible and has finite state space) then $\rho^{(K)} := c + a \sum_{i \in E} \pi_i k(i) - \sum_{i \in E} \pi_i \lambda_i \mu_i$ for some investment strategy $K = k(J)$ is the difference between the average premium income and the average payout when using the investment strategy $K = k(J)$. We refer to $\rho^{(K)}$ as the safety loading with respect to the investment strategy K .

Lemma 1. *Let $K = k(J)$ and suppose that $\rho^{(K)} \leq 0$. Then for all $u \geq 0$ it holds that $\Psi(u, K) = 1$.*

The proof of this statement is omitted since it is standard. For the remaining sections we assume that $\rho^{(K)} > 0$.

2 Fixed Investment Strategies

For a given investment strategy $K = k(J)$ our aim is to find a constant $R^{(K)} > 0$ such that for all $\varepsilon > 0$:

$$\lim_{u \rightarrow \infty} \Psi(u, K) e^{(R^{(K)} - \varepsilon)u} = 0, \text{ and } \lim_{u \rightarrow \infty} \Psi(u, K) e^{(R^{(K)} + \varepsilon)u} = \infty. \quad (2)$$

$R^{(K)}$ is then called the *adjustment coefficient* with respect to K . There are different methods available for obtaining the adjustment coefficient, see e.g. [9].

The way we pursued is similar to what Björk and Grandell [5] (see also [7]) do for the ordinary Cox model. Let the time epoch of the n^{th} entry of the environment process to state $j \in E$ be denoted by $\tau_n^{(j)}$ where $\tau_0^{(j)} \equiv 0$. We put $\tau^{(j)} := \tau_1^{(j)}$. Let $\xi_i(t) := \int_0^t \delta_{\{i\}}(J_s) ds$ be the time which J spends in some state i until time t . For $j, k \in E$ we now have to consider the function $\phi_{kj}^{(K)}$ defined by

$$\phi_{kj}^{(K)}(r) := \mathbb{E}_k \left(\exp \left(\sum_{i \in E} \left[\lambda_i h_i(r) + \frac{r^2 b^2 k(i)^2}{2} - r(c + ak(i)) \right] \xi_i(\tau^{(j)}) \right) \right).$$

where $r \geq 0$ and E_k is the expectation, given $J_0 = k$. Note that these functions are convex since the centered moment generating functions h_i are convex for all i . Using these functions we define

$$R^{(K)} := \sup \left\{ r > 0; \phi_{jj}^{(K)}(r) < 1 \forall j \in E \right\} \quad (3)$$

and get the following result which follows directly from Theorem 6 in [4].

Lemma 2. *If $R^{(K)}$ defined by (3) exists then for any $r < R^{(K)}$ we have*

$$\Psi(u, K) \leq C(K, r) e^{-ru}$$

with $C(K, r) < \infty$ for all $u \geq 0$.

In particular we have for the constant investment strategy $K_t^{(r)} \equiv \frac{a}{rb^2}$ with $r > 0$ arbitrary

$$\phi_{kj}(r) := \phi_{kj}^{(K^{(r)})}(r) = \mathbb{E}_k \left(\exp \left(\sum_{i \in E} \lambda_i h_i(r) \xi_i(\tau^{(j)}) - \left(rc + \frac{a^2}{2b^2} \right) \tau^{(j)} \right) \right).$$

Analogously to the definition of $R^{(K)}$ we now let

$$R := \sup \left\{ r > 0; \phi_{jj}(r) < 1 \forall j \in E \right\} \quad (4)$$

and recognize that $R^{(K)} \leq R$ for every $K = k(J)$ since $\phi_{jj}(r) \leq \phi_{jj}^{(K)}(r)$ for all $r \geq 0$ and $j \in E$ (which can be seen by completion of squares). It can also be shown that R has an alternative representation via the spectral radius of a certain matrix (for details see [8]).

3 The Optimal Investment Strategy

We want to find a constant $R > 0$ and an investment strategy $\hat{K} = \hat{k}(J)$ such that for all $K = k(J)$ and all $\varepsilon > 0$:

$$\lim_{u \rightarrow \infty} \Psi(u, \hat{K}) e^{(R-\varepsilon)u} = 0, \text{ and } \lim_{u \rightarrow \infty} \Psi(u, K) e^{(R+\varepsilon)u} = \infty. \quad (5)$$

R is then called the optimal adjustment coefficient of our model and $\hat{K}$ the associated optimal investment strategy. In [10] the author considers a Markov-modulated risk reserve process which is perturbed by diffusion. Using his results we obtain (for a detailed derivation see [8]):

Theorem 1. *If R defined by (4) exists, then using the constant investment strategy $K^{(R)}$ defined by $K_t^{(R)} \equiv \frac{a}{Rb^2}$ we obtain for all $u \geq 0$ and with $C < \infty$*

$$\Psi(u, K^{(R)}) \leq C e^{-Ru}$$

Finally we can show that R is indeed the optimal adjustment coefficient in the following way.

Theorem 2. *Suppose that R defined by (4) exists and consider any fixed investment strategy $K = k(J)$. For this investment strategy K we furthermore assume that $R^{(K)}$ exists and that we can find a constant $\delta > 0$ such that $\phi_{jj}^{(K)}(R^{(K)} + \delta) < \infty$ for some $j \in E$. We then have*

$$\lim_{u \rightarrow \infty} \Psi(u, K) e^{ru} = \infty$$

for all $r > R^{(K)}$ and thus in particular for all $r > R$.

Proof. Since $\phi_{jj}^{(K)}$ is convex and therefore continuous on the interior of its domain it follows that $\phi_{jj}^{(K)}$ is finite in the δ -neighborhood of $R^{(K)}$ with $\phi_{jj}^{(K)}(R^{(K)}) = 1$. Let us now assume that $J_0 = j$. It is then easy to verify that $(Y_{\tau_n^{(j)}}(0, K))_{n \in \mathbb{N}}$ is a random walk and that $\mathbb{E}_j(e^{-R^{(K)} Y_{\tau_n^{(j)}}(0, K)}) = \phi_{jj}^{(K)}(R^{(K)}) = 1$. Conditioned under $J_0 = j$ we now define the ruin probability of the shifted random walk $(Y_{\tau_n^{(j)}}(u, K))_{n \in \mathbb{N}}$ by

$$\Psi_j^{rw}(u, K) := \mathbb{P}_j\left(\inf_{n \in \mathbb{N}} Y_{\tau_n^{(j)}}(u, K) < 0\right).$$

It is obvious that $\Psi_j^{rw}(u, K) \leq \Psi_j(u, K) = \mathbb{P}_j\left(\inf_{t \geq 0} Y_t(u, K) < 0\right)$. Thus

$$\lim_{u \rightarrow \infty} \frac{\Psi_j(u, K)}{e^{-ru}} \geq \lim_{u \rightarrow \infty} \frac{\Psi_j^{rw}(u, K)}{e^{-R^{(K)}u}} e^{(r-R^{(K)})u} = \infty.$$

for all $r > R^{(K)}$ and therefore in particular for all $r > R$ since $R \geq R^{(K)}$. The limit follows from Theorem 6.5.7 and the associated remark in [9] because the distribution of $Y_{\tau_n^{(j)}}(0, K)$, i.e. the distribution of the generic random variable for the steps of the random walk, is clearly non-lattice and the existence of $R^{(K)}$ moreover implies that the safety loading of the model is strictly positive. The fact that $\Psi(u, K) = \sum_{j \in E} \Psi_j(u, K) \mathbb{P}(J_0 = j)$ and the assumption $\mathbb{P}(J_0 = j) > 0$ for all $j \in E$ implies the result.

4 A Comparison with the Compound Poisson Model

It is intuitively clear that we can relate a compound Poisson model to the Markov-modulated Poisson model in a natural way by averaging over the environment (cf. [2]). More precisely, we consider a compound Poisson model with investment where the intensity of the claim arrival process and respectively the claim size distribution are defined by

$$\lambda^* = \sum_{i \in E} \pi_i \lambda_i \quad \text{and} \quad B^* = \sum_{i \in E} \frac{\pi_i \lambda_i}{\lambda^*} B_i.$$

We refer to this model as the related compound Poisson model. Note, that its claims have exponential moments since

$$h^*(r) := \int_0^\infty e^{rx} dB^*(x) - 1 = \left(\sum_{i \in E} \frac{\pi_i \lambda_i}{\lambda^*} \int_0^\infty e^{rx} dB_i(x) \right) - 1 = \sum_{i \in E} \frac{\pi_i \lambda_i}{\lambda^*} h_i(r).$$

It is shown in [6] that the optimal adjustment coefficient R^* of the related compound Poisson model with investment is given as the strictly positive solution of the equation $\lambda^* h^*(r) = rc + \frac{a^2}{2b^2}$ and that the corresponding optimal investment strategy is $K^{(R^*)}$.

Theorem 3.

(i) Let R and R^* be the adjustment coefficients of the Markov-modulated Poisson model and its related compound Poisson model under optimal investment respectively as defined above. Then,

$$R \leq R^*.$$

(ii) The corresponding optimal investment strategies $K^{(R)}$ and $K^{(R^*)}$ respectively satisfy

$$|K_t^{(R)}| \geq |K_t^{(R^*)}|.$$

Proof. (i) Consider any fixed $r > R^*$ and recall from the theory of Markov chains that $\mathbb{E}_j(\xi_i(\tau^{(j)})) = \pi_i \mathbb{E}_j(\tau^{(j)})$ for all $i \in E$. Using Jensen's inequality we thus get

$$\phi_{jj}(r) \geq \exp \left(\left(\sum_{i \in E} \pi_i \lambda_i h_i(r) - \left(rc + \frac{a^2}{2b^2} \right) \right) \mathbb{E}_j(\tau^{(j)}) \right).$$

Moreover, we have $\lambda^* h^*(r) = \sum_{i \in E} \pi_i \lambda_i h_i(r)$. Since R^* solves the equation $\lambda^* h^*(r) = rc + \frac{a^2}{2b^2}$ it consequently follows that

$$\sum_{i \in E} \pi_i \lambda_i h_i(r) - \left(rc + \frac{a^2}{2b^2} \right) = \lambda^* h^*(r) - \left(rc + \frac{a^2}{2b^2} \right) \geq 0$$

which implies $\phi_{jj}(r) \geq 1$ and consequently $R \leq R^*$.

(ii) The assertion follows directly from part (i).

Note that in the case where the drift parameter a is equal to zero the investment strategy $K^{(r)}$ provides not to invest into the portfolio for all $r > 0$. In this case Theorem 3 thus coincides with Theorem 3 in [3] where it is shown that the adjustment coefficient of the Markov-modulated Poisson model without investment does not exceed the adjustment coefficient of its related compound Poisson model without investment.

References

1. Asmussen S (1989) Risk theory in a Markovian environment. *Scandinavian Actuarial Journal* 2:69–100.
2. Asmussen S (2000) *Ruin probabilities*. World Scientific, Singapore.
3. Asmussen S, O’Cinneide C (2002) On the tail of the waiting time in a Markov-modulated $M/G/1$ queue. *Operations Research* 50:559–565.
4. Bäuerle N, Kötter M (2006) Markov-modulated Diffusion Risk Models. Preprint.
5. Björk T, Grandell J (1988) Exponential inequalities for ruin probabilities in the Cox case. *Scandinavian Actuarial Journal* 1-2:77–111.
6. Gaier J, Grandits P, Schachermayer W (2003) Asymptotic ruin probabilities and optimal investment. *The Annals of Applied Probability* 13:1054–1076.
7. Grandell J (1991) *Aspects of Risk Theory*. Springer, New York.
8. Kötter M. (2006) *Optimal investment in time inhomogeneous Poisson models*. PhD Thesis, University of Hannover, Hannover.
9. Rolski T, Schmidli H, Schmidt V, Teugels J (1999) *Stochastic Processes for Insurance and Finance*. John Wiley & Sons, Chichester.
10. Schmidli H (1995) Cramér-Lundberg approximations for ruin probabilities of risk processes perturbed by diffusion. *Insurance: Mathematics & Economics*. 16:135–149.

Optimal Portfolios Under Bounded Shortfall Risk and Partial Information

Ralf Wunderlich¹, Jörn Sass² and Abdelali Gabih³

¹ Mathematics Group, Zwickau University of Applied Sciences, Germany
ralf.wunderlich@fh-zwickau.de

² RICAM, Austrian Academy of Sciences, Linz, Austria
joern.sass@oeaw.ac.at

³ Dept. of Mathematics and Computer Science, Leipzig University, Germany,
abdelali.gabih@math.uni-leipzig.de

1 Introduction

This paper considers the optimal selection of portfolios for utility maximizing investors under a shortfall risk constraint for a financial market model with partial information on the drift parameter. It is known that without risk constraint the distribution of the optimal terminal wealth often is quite skew. In spite of its maximum expected utility there are high probabilities for values of the terminal wealth falling short a prescribed benchmark. This is an undesirable and unacceptable property e.g. from the viewpoint of a pension fund manager. While imposing a strict restriction to portfolio values above a benchmark leads to considerable decrease in the portfolio's expected utility, it seems to be reasonable to allow shortfall and to restrict only some shortfall risk measure.

A very popular risk measure is value at risk (VaR) which takes into account the probability of a shortfall but not the actual size of the loss. Therefore we use the so-called expected loss criterion resulting from averaging the magnitude of the losses. And in fact, e.g. in Basak, Shapiro [1] it is shown that the distribution of the resulting optimal terminal wealth has more desirable properties.

We use a financial market model which allows for a non-constant drift which is not directly observable. In particular, we use a hidden Markov model (HMM) where the drift follows a continuous time Markov chain. In [13] it was shown that on market data utility maximizing strategies based on such a model can outperform strategies based on the assumption of a constant drift parameter. Extending these results to portfolio optimization problems under risk constraints we obtain in Theorem 2 and 3 quite explicit representations for the form of the optimal terminal wealth and the trading strategies which can be computed using Monte Carlo methods.

For additional topics such as stochastic interest rates, stochastic volatility, motivation of the model, Malliavin calculus, aspects of parameter estimation, and for more references concerning partial information see [8, 9, 13]. For an control theoretic approach see Rieder, Bäuerle [12].

For more background, references and results on optimization under risk constraints see e.g. Basak, Shapiro [1], Gundel, Weber [7], Lakner, Nygren [11], and [3, 4, 5].

We acknowledge support from the Austrian Science Fund FWF, P 17947.

2 An HMM for the Stock Returns

Let $(\Omega, \mathcal{A}, P)$ be a complete probability space, $T > 0$ the terminal trading time, and $\mathcal{F} = (\mathcal{F}_t)_{t \in [0, T]}$ a filtration in $\mathcal{A}$ satisfying the usual conditions. We consider one *money market* with interest rates equal 0 (to simplify notation) and n stocks whose *prices* $S = (S_t)_{t \in [0, T]}$, $S_t = (S_t^1, \dots, S_t^n)^\top$ evolve according to

$$dS_t = \text{Diag}(S_t)(\mu_t dt + \sigma dW_t), \quad S_0 \in \mathbb{R}^n,$$

where $W = (W_t)_{t \in [0, T]}$ is an n -dimensional Brownian motion w.r.t. $\mathcal{F}$, and σ is the non-singular $(n \times n)$ -*volatility-matrix*. The *return process* $R = (R_t)_{t \in [0, T]}$ is defined by $dR_t = (\text{Diag}(S_t))^{-1} dS_t$. We assume that $\mu = (\mu_t)_{t \in [0, T]}$, the *drift process* of the return, is given by $\mu_t = B Y_t$, where $Y = (Y_t)_{t \in [0, T]}$ is a stationary, irreducible, *continuous time Markov chain* independent of W with state space $\{e_1, \dots, e_d\}$, the standard unit vectors in $\mathbb{R}^d$. The columns of the *state matrix* $B \in \mathbb{R}^{n \times d}$ contain the d possible states of μ_t . Further Y is characterized by its *rate matrix* $Q \in \mathbb{R}^{d \times d}$, where $\lambda_k = -Q_{kk} = \sum_{l=1, l \neq k}^d Q_{kl}$ is the rate of leaving e_k and Q_{kl}/λ_k is the probability that the chain jumps to e_l when leaving e_k .

Since the *market price of risk*, $\vartheta_t = \sigma^{-1} \mu_t = \sigma^{-1} B Y_t$, $t \in [0, T]$, is uniformly bounded the density process $(Z_t)_{t \in [0, T]}$ defined by $dZ_t = -Z_t \vartheta_t^\top dW_t$, $Z_0 = 1$, is a martingale. By $d\tilde{P} = Z_T dP$ we define the *risk-neutral* probability measure. $\tilde{\mathbb{E}}$ will denote expectation with respect to $\tilde{P}$. Girsanov's Theorem guarantees that $d\tilde{W}_t = dW_t + \vartheta_t dt$ defines a $\tilde{P}$ -Brownian motion. The definition of R yields

$$R_t = \int_0^t B Y_s ds + \sigma W_t = \sigma \tilde{W}_t, \quad t \in [0, T]. \quad (1)$$

We consider the case of *partial information* meaning that an investor can only observe the prices. Neither the drift process nor the Brownian motion are observable. Only the events of $\mathcal{F}^S$, the augmented filtration generated by S , can be observed and hence all investment decisions have to be adapted to $\mathcal{F}^S$. Note that $\mathcal{F}^S = \mathcal{F}^R = \mathcal{F}^{\tilde{W}}$.

A *trading strategy* $\pi = (\pi_t)_{t \in [0, T]}$ is an n -dimensional $\mathcal{F}^S$ -adapted, measurable process which satisfies $\int_0^T \|\pi_t\|^2 dt < \infty$. The wealth invested in the i -th stock at time t is π_t^i and $X_t^\pi - \mathbf{1}_n^\top \pi_t$ is invested in the money market, where $(X_t^\pi)_{t \in [0, T]}$ is the corresponding *wealth process*. For *initial capital* $x_0 > 0$ it is defined by $dX_t^\pi = \pi_t^\top (\mu_t dt + \sigma dW_t)$, $X_0^\pi = x_0$. A trading strategy π is called *admissible* if $P(X_t^\pi \geq 0 \text{ for all } t \in [0, T]) = 1$. By Itô's rule

$$X_t^\pi = x_0 + \int_0^t \pi_s^\top \sigma d\tilde{W}_s, \quad t \in [0, T]. \quad (2)$$

A *utility function* $U : [0, \infty) \rightarrow \mathbb{R} \cup \{-\infty\}$ is strictly increasing, strictly concave, twice continuously differentiable, and satisfies the Inada conditions $\lim_{x \rightarrow \infty} U'(x) = 0$,

$\lim_{x \rightarrow 0} U'(x) = \infty$. Moreover, we impose as a technical condition that the function $I = (U')^{-1}$ satisfies $E[Z_T I(yZ_T)] < \infty$ and $E[Z_T^2 | I'(yZ_T)] < \infty$ for all $y > 0$.

We want to maximize the expected utility of the terminal wealth X_T^π but also constrain the risk that the terminal wealth falls short of a benchmark $q > 0$. The shortfall risk is measured in terms of the expected loss, which is computed by averaging the loss $(X_T^\pi - q)^-$ w.r.t. the risk neutral measure $\tilde{P}$.

For given $x_0 > 0$, $q > 0$ and $\varepsilon > 0$, the bound for the shortfall risk, we thus want to

$$\text{maximize } E[U(X_T^\pi)] \quad \text{s.t.} \quad \tilde{E}[(X_T^\pi - q)^-] \leq \varepsilon$$

over all admissible trading strategies π . We will denote an optimal strategy by π^* and the corresponding wealth process by X^* . The expected loss corresponds to the price of a derivative to hedge against the shortfall. So by paying ε now, one can hedge against the risk to fall short of q . In [5] we also consider a generalized expected loss criterion where the measure $\tilde{P}$ used for the averaging of the loss $(X_T^\pi - q)^-$ is replaced by some arbitrary measure equivalent to P .

To determine optimal trading strategies we have to find a good estimator for the drift process. By (1) we are in the classical situation of *HMM filtering* with signal Y and observation R , where we want to determine the filter $E[Y_t | \mathcal{F}_t^R] = E[Y_t | \mathcal{F}_t^S]$ for Y_t . By Theorem 4 in [4], Bayes' Law, and using $\mathbf{1}_d^\top Y_t = 1$ we get

Theorem 1. *The filter $\eta_t = E[Y_t | \mathcal{F}_t^S]$ (for Y), the unnormalized filter $\mathcal{E}_t = \tilde{E}[Z_T^{-1} Y_t | \mathcal{F}_t^S]$ (for Y) and the conditional density $\zeta_t = E[Z_t | \mathcal{F}_t^S]$ (filter for Z_t) satisfy $\eta_t = \zeta_t \mathcal{E}_t$, $\zeta_t^{-1} = \mathbf{1}_d^\top \mathcal{E}_t$, and*

$$\mathcal{E}_t = E[Y_0] + \int_0^t Q^\top \mathcal{E}_s ds + \int_0^t \text{Diag}(\mathcal{E}_s) B^\top (\sigma \sigma^\top)^{-1} dR_s, \quad t \in [0, T].$$

Moreover, $\zeta_t^{-1} = \tilde{E}[Z_t^{-1} | \mathcal{F}_t^S]$ and $\zeta_t^{-1} = 1 + \int_0^t (B \mathcal{E}_s)^\top (\sigma \sigma^\top)^{-1} dR_s$, $t \in [0, T]$.

3 Optimal Trading Strategies

One has to take care about selecting the bound ε for the shortfall risk. Choosing a value which is too small, there is no admissible solution of the problem, because the risk constraint cannot be satisfied. In [5] we find that the risk constraint is binding for $\varepsilon \in (\underline{\varepsilon}, \bar{\varepsilon})$ where $\underline{\varepsilon} = \max(0, q - x_0)$ and $\bar{\varepsilon}$ is the expected loss of the optimal terminal wealth of the Merton problem, i.e. the optimization problem without risk constraint.

The dynamic portfolio optimization problem can be splitted into two problems - the static and the representation problem. While the static problem is concerned with the form of the optimal terminal wealth the representation problem consists in the computation of the optimal trading strategy.

The next Theorem gives the form of the optimal terminal wealth. The proof which can be found in [5, Section 5] adopts the common convex-duality approach by introducing the convex conjugate of the utility function U with an additional term capturing the risk constraint.

Theorem 2. Let $\varepsilon \in (\underline{\varepsilon}, \bar{\varepsilon})$, then the optimal terminal wealth is

$$X_T^* = f(\zeta_T) = f(\zeta_T; y_1^*, y_2^*) := \begin{cases} I(y_1^* \zeta_T) & \text{for } \zeta_T \in (0, \underline{z}] \\ q & \text{for } \zeta_T \in (\underline{z}, \bar{z}] \\ I((y_1^* - y_2^*) \zeta_T) & \text{for } \zeta_T \in (\bar{z}, \infty). \end{cases}$$

where $I = (U')^{-1}$, $\underline{z} = \frac{U'(q)}{y_1^*}$ and $\bar{z} = \frac{U'(q)}{y_1^* - y_2^*}$.

The real numbers $y_1^*, y_2^* > 0$ solve the equations

$$\tilde{\mathbb{E}}(f(\zeta_T; y_1, y_2)) = x_0 \quad \text{and} \quad \tilde{\mathbb{E}}(f(\zeta_T; y_1, y_2) - q)^- = \varepsilon.$$

The solution of the above system of equations exists and is unique.

If $\mathbb{E}[|U(X_T^*)|] < \infty$ then X_T^* is the P -almost sure unique solution.

For the solution of the representation problem we can proceed as in [13, Section 4] to find the optimal trading strategy. Since $f(z)$ is not differentiable at $\underline{z}$ and $\bar{z}$ we have to use some approximation arguments to show that the chain rule for the Malliavin derivative yields $D_t f(\zeta_T) = f'(\zeta_T) D_t \zeta_T$, where f' is defined piecewise on the intervals $(0, \underline{z})$, $(\underline{z}, \bar{z})$, $(\bar{z}, \infty)$. The proof of the following theorem can be found in [5, Section 7]. It uses the Martingale Representation Theorem and Clark's Formula in $\mathcal{M}_{1,1}$, see Karatzas, Ocone, Li [10].

Theorem 3. Let $X_T^* \in L^2(\tilde{P})$ and $I'(y \zeta_T) \in L^p(\tilde{P})$ for all $y > 0$ and some $p > 1$. Then for the optimal strategy it holds

$$\begin{aligned} \pi_t^* &= (\sigma \sigma^\top)^{-1} \tilde{\mathbb{E}}[f'(\zeta_T) \sigma D_t \zeta_T \mid \mathcal{F}_t^S], \quad \text{where} \\ \sigma D_t \zeta_T &= -\zeta_T^2 \left(B \mathcal{E}_t + \int_t^T (\sigma D_t \mathcal{E}_s) B^\top (\sigma \sigma^\top)^{-1} dR_s \right), \\ \sigma D_t \mathcal{E}_s &= B \text{Diag}(\mathcal{E}_t) + \int_t^s (\sigma D_t \mathcal{E}_u) Q du + \int_t^s (\sigma D_t \mathcal{E}_u) \text{Diag} \left(B^\top (\sigma \sigma^\top)^{-1} dR_u \right). \end{aligned}$$

4 Numerical Example

We consider a financial market with $n = 1$ stock with $d = 2$ states of the drift $b_1 = 0.74$ and $b_2 = -0.36$, volatility $\sigma = 0.25$, rates $\ell_1 = 15$, $\ell_2 = 10$, horizon $T = 1$, initial capital $x_0 = 1$, benchmark $q = 0.9$ and bound $\varepsilon = 0.05$. i.e. after one year we would like to have still 90% of our initial wealth with at most an expected loss of 0.05. From 10 million simulated paths we estimate for logarithmic utility the parameters y_1 and y_2 and estimate the expectations of the terminal wealth, its utility and the shortfall risk. The results are given in Table 1, where we compare the above values with the pure stock portfolio and the optimal portfolio for the Merton problem which contains no risk constraint.

The expected loss of the pure stock strategy nearly coincides with the prescribed bound $\varepsilon = 0.05$ while the Merton strategy exhibits a much larger risk measure of nearly 4ε . The constrained strategy has a slightly lower expected utility as well as expected terminal wealth than the unconstrained optimal strategy while it clearly outperforms the pure stock investment. Figure 1 shows the estimated probability

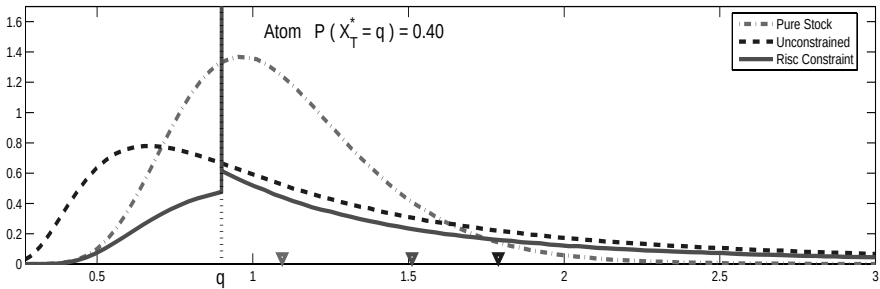


Fig. 1. Distribution of terminal wealth

Table 1. Estimated expectations of terminal wealth, its utility and expected loss

	exp. term. wealth $E[X_T]$	expected utility $E[U(X_T)]$	expected shortfall $\tilde{E}[(X_T - q)^-]$
pure stock	1.095	0.049	0.053
unconstrained (Merton)	1.789	0.227	0.182
with risk constraint	1.512	0.186	0.050 = ε

density functions of the terminal wealth considered above. Additionally, on the horizontal axes the expected terminal wealth $E[X_T]$ for the considered portfolios are marked.

The figure clearly shows the reason for the large expected loss of the Merton portfolio. There is a large probability for values in the “shortfall region” $[0, q)$. On the other hand there are considerable tail probabilities leading to the high expectation of the terminal wealth as well as its utility. For the constrained portfolio probability mass from that “shortfall region” but also from the tail is shifted to build up the atom at the point q (benchmark). This results in a slightly smaller expectations of the terminal wealth and its utility. According to Theorem ?? the atom at q has the size $P(X_T^* = q) = P(\underline{z} < \zeta_T \leq \bar{z})$. In the density plot it is marked by a vertical line at q .

The optimal strategy can be evaluated using the representation given in Theorem 3 by an Monte-Carlo approximation of the conditional expectation from a sufficiently large number of paths of ζ_t and the Malliavin derivative $D_t \zeta_T$. These paths can be obtained via numerical solution of the SDEs given in Theorem 1 and 3. A more detailed analysis of the strategies and application to market data will be the topic of forthcoming papers.

References

1. Basak S, Shapiro A (2001) Value-at-risk based risk management: Optimal policies and asset prices. The Review of Financial Studies 14: 371–405

2. Elliott RJ (1993) New finite-dimensional filters and smoothers for noisily observed Markov chains. *IEEE Transactions on Information Theory* 39: 265–271
3. Gabih A, Grecksch W, Wunderlich R (2005) Dynamic portfolio optimization with bounded shortfall risks. *Stochastic Analysis and Applications* 23: 579–594
4. Gabih A, Grecksch W, Richter M, Wunderlich R (2006) Optimal portfolio strategies benchmarking the stock market. *Mathematical Methods of Operations Research*, to appear
5. Gabih A, Sass J, Wunderlich R (2006) Utility maximization under bounded expected loss. *RICAM report 2006-24*
6. Gandy R (2005) Portfolio optimization with risk constraints. PhD-Dissertation, Universität Ulm
7. Gundel A, Weber S (2005) Utility maximization under a shortfall risk constraint. Preprint
8. Hahn M, Putschögl W, Sass J (2006) Portfolio Optimization with Non-Constant Volatility and Partial Information. *Brazilian Journal of Probability and Statistics*, to appear
9. Haussmann U, Sass J (2004) Optimal terminal wealth under partial information for HMM stock returns. In: G. Yin and Q. Zhang (eds.): *Mathematics of Finance: Proceedings of an AMS-IMS-SIAM Summer Conference June 22-26, 2003, Utah*, *AMS Contemporary Mathematics* 351: 171–185
10. Karatzas I, Ocone DL, and Li J (1991) An extension of Clark’s formula. *Stochastics and Stochastics Reports* 37: 127–131
11. Lakner P, Nygren LM (2006) Portfolio optimization with downside constraints. *Mathematical Finance* 16: 283–299
12. Rieder U, Bäuerle N (2005) Portfolio optimization with unobservable Markov-modulated drift process. *J. Appl. Prob.* 43: 362–378
13. Sass J, Haussmann UG (2004) Optimizing the terminal wealth under partial information: The drift process as a continuous time Markov chain. *Finance and Stochastics* 8: 553–577

OR for Simulation and Its Optimization

Nico M. van Dijk and Erik van der Sluis

Fac. of Economics and Business, University of Amsterdam, The Netherlands

`N.M.vanDijk@uva.nl`

`H.J.vanderSluis@uva.nl`

Summary. This is an expository paper to promote the potential of OR (Operations Research) for simulation. Three applications will therefore be presented which include call centers, check-in at airports, and performance bounds for production lines. The results indicate that (classical and new) OR results might still be most fruitful if not necessary for practical simulation and its optimization.

1 Introduction

Simulation or more precisely as meant in the setting of this paper: discrete event simulation is known as a most powerful tool for the evaluation of logistical systems such as arising in manufacturing, communications or the service industry (banks, call centers, hospitals). A general characterization is that these systems:

- are complex
- involve stochastics and
- require some form of improvement
(such as by infrastructure, lay-out or work procedures).

Analytic methods, such as standardly covered by the field of OR (Operations Research), in contrast, only apply if:

- the systems are sufficiently simple and
- special assumptions are made on the stochastics involved.

On the other hand, also simulation has a number of limitations:

1. The lack of insights;
2. Techniques for optimization;
3. The confidence that can be adhered to the results.

In this paper, therefore, it will be illustrated how the discipline of OR (Operations Research) might still be beneficial for each of these three aspects.

First in section 2, it is shown that a classic queueing insight in combination with simulation already lends to counterintuitive results and ways for optimization for the simple question of whether we should pool servers or not such as in call centers. Next, in section 3, a check-in problem is presented. First, it is argued that simulation is necessarily required. Next OR-technique, in this case LP-formulation, is used for optimization. Finally, in section 4, it is shown that analytic queueing results can still be most supportive for the simulation of queueing network such as production lines.

The results in this paper all rely upon earlier and practical research that has been reported in more extended and specific technical papers for each of the applications separately (cf. [1], [2], [3]) but without the common message of OR for simulation as promoted in this paper.

2 To Pool or Not in Call Centers

Should we pool servers or not? This seems a simple question of practical interest, such as for counters in postal offices, check-in desks at airports, physicians within hospitals, up to agent groups within or between call centers. The general perception seems to exist that pooling capacities is always advantageous.

An Instructive Example (Queueing)

This perception seems supported by the standard delay formula for a single (exponential) server with arrival rate λ and service rate μ : $D = 1/(\mu - \lambda)$. Pooling two servers thus seems to reduce the mean delay by roughly a factor 2 according to $D = 1/(2\mu - 2\lambda)$.

However, when different services are involved in contrast, a second basic result from queueing theory is to be realized: Pollaczek-Khintchine formula. This formula, which is exact for the single server case, expresses the effect of service variability, by:

$$\begin{aligned} \mathbf{W}_G &= \frac{1}{2}(1 + c^2)\mathbf{W}_E \text{ with } c^2 = \sigma^2/\tau^2 \text{ and} \\ \mathbf{W}_G &\quad \text{the mean waiting time under a general (and } E \text{ for exponential)} \\ &\quad \text{service distribution with mean } \tau \text{ and standard deviation } \sigma. \end{aligned}$$

By mixing different services (call types) extra service variability is brought in which may lead to an increase of the mean waiting time.

This is illustrated in the figure below for the situation of two job (call) types 1 and 2 with mean service (call) durations $\tau_1 = 1$ and $\tau_2 = 10$ minutes but arrival rates $\lambda_1 = 10\lambda_2$. The results show that the unpooled case is still superior, at least for the average waiting time $\mathbf{W}_A$. Based on these queueing insights, a two-way or one-way overflow scenario can now be suggested, which leads to further improvement as also illustrated in Fig. 1.

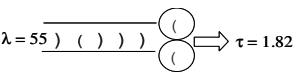
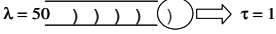
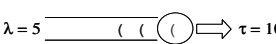
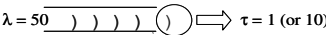
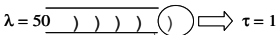
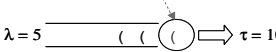
Pooled system	$W_A = 6.15$	Unpooled system	$W_A = 4.55$
$\lambda = 55$ 	$W = 6.15$	$\lambda = 50$  $\lambda = 5$ 	$W_1 = 2.50$ $W_2 = 25.0$
Two-way overflow	$W_A = 4.11$	One-way overflow	$W_A = 3.92$
$\lambda = 50$  $\lambda = 5$ 	$W_1 = 3.66$ $W_2 = 8.58$	$\lambda = 50$  $\lambda = 5$ 	$W_1 = 1.80$ $W_2 = 25.2$

Fig. 1. Pooling scenarios

A Combined Approach

To achieve these improvements simulation is necessarily required. A combination of queueing for its insights to suggest scenarios and of simulation for evaluating these scenarios thus turns out to be fruitful.

Call Centers (Large Number of Servers)

Similar results can also obtained for larger number of servers, say with 10, 50 or 100 servers, such as arising in realistic call centers. This is illustrated in Table 1. The one-way overflow scenario turns out to be superior to both the pooled and the unpooled scenario for realistically large numbers of call centers agents. (Here the mix ratio of short and long services is similar as in the example above. For further details, see [2]).

Table 1. Results for two server groups each with s servers

s	Pooled	Unpooled			One-way overflow			% overflow
	W_P	W_1	W_2	W_A	W_1	W_2	W_A	
1	11.53	4.49	45.24	8.18	3.40	45.47	7.20	5.1%
5	1.76	0.78	7.57	1.41	0.53	7.75	1.20	4.9%
10	0.71	0.34	3.52	0.63	0.21	3.66	0.52	4.5%
20	0.26	0.15	1.44	0.26	0.08	1.54	0.21	4.0%
50	0.05	0.04	0.38	0.07	0.02	0.42	0.05	2.9%

3 Check-In Planning

Problem Formulation and Combined Approach

Check-in desks and desk-labor hours can be a scarce resource at airports. To minimize the required number of desks two essentially different optimization problems

are involved:

- P1: A minimization of the required number of desks for a given flight.
- P2: A minimization and scheduling for all flights during a day.

A two-step procedure is therefore proposed:

- Step 1: For P1 as based upon simulation
- Step 2: For P2 as based upon LP (Linear Programming)

Step 1: Simulation

As the check-in process is highly transient (fixed opening interval, non-homogeneous arrivals during opening hours and initial bias at opening time) transient (or terminating) simulation will necessarily be required.

In step 1 therefore the required number of desks will have to be determined by terminating simulation for each hour of the day and separate (group of) flights.

Step 2: OR (Linear Programming)

Next in step 2 the desks are to be scheduled for the different flights so as to minimize the total number of desks and desk-(labor)-hours. Here additional practical conditions may have to be taken into account such as most naturally that desks for one and the same flight should be adjacent.

Example

As a simple (fictitious) example consider the desk requirements for 5 flights during 9 hours (periods), as determined by step 1. The total number of desks required then never exceeds 4.

However, a straightforward Earliest Release Date (ERD) desk allocation as shown in the figure below would lead to an unfeasible solution, as the desks for flight 5 are not adjacent. (This could be resolved by using two more desks 5, 6 and assigning desks 4, 5 and 6 to flight 5). However, in this example a feasible solution with 4 desks is easily found.

Infeasible schedule									
d\i	1	2	3	4	5	6	7	8	9
1	1	1	1	3	3	3	5	5	5
2	1	1	1	3	3	3	5	5	5
3	1	1	1		4	4	4		
4				2	2	2	5	5	5

Feasible schedule									
d\i	1	2	3	4	5	6	7	8	9
1	1	1	1		4	4	4		
2	1	1	1	3	3	3	5	5	5
3	1	1	1	3	3	3	5	5	5
4				2	2	2	5	5	5

Fig. 2. Example of an infeasible and a feasible schedule

As shown in [3] also for more realistic orders with hundreds of flights an optimal solution can be found by solving an LP-formulation as given below. The combination of (terminating) simulation and LP-optimization so turned out to be most beneficial.

$$\begin{array}{ll}
 \min D & D : \text{Total number of desks required;} \\
 \text{s.t. } n_f \leq d_f \leq D & \forall f \quad d_f : \text{Largest desk number to flight } f; \\
 d_f + n_g \leq d_g & \text{or} \quad n_f : \text{Number of desks for flight } f. \\
 d_g + n_f \leq d_f & \forall f, g
 \end{array}$$

4 Product-Form Bounds for Production Lines

Analytic results for queueing networks, most notably so-called product forms, are generally violated by practical aspects such as finite capacities. Nevertheless, such results might still lead to analytic performance bounds.

These bounds, in turn, can be most useful for simulation as for:

- verification and validation purposes and
- determining orders of magnitude.

As an example, consider the simple but generic assembly line structure as shown with 4 service stations, numbered 1, $\dots$, 4 and finite capacity constraints T_1 for the total number of jobs at stations 1 and 2 (cluster1) and T_2 at stations 3 and 4 (cluster 2).

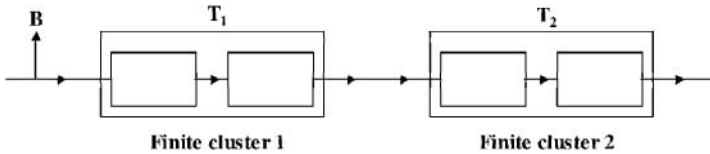


Fig. 3. Assembly line with two finite buffers

The system has an arrival rate of λ jobs per unit of time. Station i has a service rate $\mu_i(k)$ when k jobs are present. Let n_i denote the number of jobs at station i , $i = 1, \dots, 4$ and t_j the total number of jobs at cluster j , $j = 1, 2$. ($t_1 = n_1 + n_2$ and $t_2 = n_3 + n_4$). When the first cluster is saturated ($t_1 = T_1$), an arriving job is lost. When the second cluster is saturated ($t_2 = T_2$), the service at cluster 1 (that is at both stations) is stopped.

As simple as the system may look to analyze, there is no simple expression for the the loss probability B or *throughput* $H = \lambda(1 - B)$.

As outlined in [1], to enforce an analytic closed product-form expression the following modification is therefore suggested:

- When cluster 2 is saturated ($t_2 = T_2$), *stop* the input and *both stations* at cluster 1.
- When cluster 1 is saturated ($t_1 = T_1$), *stop* cluster 2, that is, *both stations* at cluster 2.

Indeed, with $\mathbf{n} = (n_1, n_2, n_3, n_4)$, e_i the unit vector for the i^{th} component and $\mathbf{1}_A$ the indicator of event A , under the above modification one easily verifies the station balance equations at S_U the set of admissible states:

$$S_U = \{\mathbf{n} | t_l = n_1 + n_2 \leq T_1; t_2 = n_3 + n_4 \leq T_2; t_1 + t_2 \neq T_1 + T_2\}$$

as

$$\left\{ \begin{array}{l} \pi(\mathbf{n})\mu_1(n_1)\mathbf{1}_{(t_2 < T_2)} = \pi(\mathbf{n} - e_1)\lambda\mathbf{1}_{(t_2 < T_2)} \\ \pi(\mathbf{n})\mu_2(n_2)\mathbf{1}_{(t_2 < T_2)} = \pi(\mathbf{n} - e_2 + e_1)\mu_1(n_1 + 1)\mathbf{1}_{(t_2 < T_2)} \\ \pi(\mathbf{n})\mu_3(n_1)\mathbf{1}_{(t_1 < T_1)} = \pi(\mathbf{n} - e_3 + e_2)\mu_2(n_2 + 1)\mathbf{1}_{(t_1 < T_1)} \\ \pi(\mathbf{n})\mu_4(n_1)\mathbf{1}_{(t_1 < T_1)} = \pi(\mathbf{n} - e_4 + e_3)\mu_3(n_3 + 1)\mathbf{1}_{(t_1 < T_1)} \\ \pi(\mathbf{n})\lambda\mathbf{1}_{(t_1 < T_1)} = \pi(\mathbf{n} + e_4)\mu_4(n_4 + 1)\mathbf{1}_{(t_1 < T_1)} \end{array} \right\}$$

by the product-form:

$$\pi(\mathbf{n})c\lambda^{n_1+n_2+n_3+n_4}\prod_{i=1}^4\left[\prod_{k=1}^{n_i}\mu_i(k)\right]^{-1}$$

for any $\mathbf{n} \in S_U$ and with c a normalizing constant. Clearly, by the modification, a simple upper bound $\mathbf{B}_U \geq \mathbf{B}$ is obtained. Conversely, also a lower bound product-form modification $\mathbf{B}_L$ can be suggested. Here

$$\mathbf{B}_U = \sum_{\{n|t_1=T_1 \text{ or } t_2=T_2\}} \pi_U(\mathbf{n})$$

Below some numerical results are given for the case of single server stations with μ_i the service speed of station i . $\mathbf{B}_L$ and $\mathbf{B}_U$ are the easily obtained lower and upper bound for the blocking probability, $\mathbf{B}_{av} = \frac{(\mathbf{B}_L + \mathbf{B}_U)}{2}$ and $\mathbf{B}$ is obtained by numerical computation.

Table 2. Lower and upper bounds of the loss probability $\mathbf{B}$ (and throughput $\mathbf{H}$ by $\mathbf{H} = \lambda(1-\mathbf{B})$) for a finite two-cluster tandem example

μ_1	μ_2	μ_3	μ_4	T_1	T_2	$\mathbf{B}_L$	$\mathbf{B}_U$	$\mathbf{B}_{av}$	$\mathbf{B}$
1	1	1	1	3	5	0.33	0.52	0.43	0.42
1	1	1	1	6	6	0.25	0.40	0.33	0.30
1	1	1	1	8	8	0.20	0.33	0.27	0.24
2	2	1	1	10	10	0.10	0.17	0.14	0.12
1	2	3	2	10	10	0.054	0.101	0.078	0.084
1.1	2	3	2	10	10	0.021	0.065	0.048	0.049

References

1. Van Dijk, N. M., Van der Sluis, E. (2002). Simple Product-Form Bounds for Queueing Networks with Finite Clusters. Annals of Operations Research, 113, 175-195.

2. Van Dijk, N. M., Van der Sluis, E. (2006). To Pool or not to Pool in Call Centers. To appear in Production and Operations Management.

3. Van Dijk, N. M., Van der Sluis, E. (2006). Check-in computation and optimization by simulation and IP in combination. European Journal of Operations Research 171, 1152-1168.

Stochastic Programming

Multistage Stochastic Programming Problems; Stability and Approximation

Vlasta Kanková

Institute of Information Theory and Automation, Academy of Sciences of the
Czech Republic, Praha, Czech Republic
kankova@utia.cas.cz

1 Introduction

A multistage stochastic programming problem can be introduced as a finite system of parametric one-stage optimization problems with an inner type of dependence and mathematical (mostly conditional) expectation in objective functions of the individual problems (for more details see e.g. [1], [3], [8]). The constraints sets can depend on the “underlying” probability measure.

To recall the multistage (M -stage, $M \geq 2$) stochastic programming problem let $\xi^j := \xi^j(\omega)$, $j = 1, \dots, M$ denote an s -dimensional random vector defined on a probability space $(\Omega, \mathcal{S}, P)$; $F^{\xi^j}(z^j)$, $z^j \in R^s$, $j = 1, 2, \dots, M$ denote the distribution function of the ξ^j and $F^{\xi^k|\bar{\xi}^{k-1}}(z^k|\bar{z}^{k-1})$, $z^k \in R^s$, $\bar{z}^{k-1} \in R^{(k-1)s}$, $k = 2, \dots, M$ denote the conditional distribution function (ξ^k conditioned by $\bar{\xi}^{k-1}$); $Z_{F^{\xi^j}} \subset R^s$, $j = 1, 2, \dots, M$ denote the support of the probability measure corresponding to F^{ξ^j} . Let, moreover, $g_0^M(\bar{x}^M, \bar{z}^M)$ be a continuous function defined on $R^{nM} \times R^{sM}$; $X^k \subset R^n$, $k = 1, 2, \dots, M$ be a nonempty set; $\mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k) := \mathcal{K}_{F^{\xi^{k+1}|\bar{\xi}^k}}^{k+1}(\bar{x}^k, \bar{z}^k)$, $k = 1, \dots, M-1$ be a multifunction mapping $R^{nk} \times R^{sk}$ into the space of subsets of R^n . $\bar{\xi}^k := \bar{\xi}^k(\omega) = [\xi^1, \dots, \xi^k]$; $\bar{z}^k = [z^1, \dots, z^k]$, $z^j \in R^s$; $\bar{x}^k = [x^1, \dots, x^k]$, $x^j \in R^n$; $\bar{X}^k = X^1 \times X^2 \times \dots \times X^k$; $\bar{Z}_{\mathcal{F}}^k = Z_{F^{\xi^1}} \times Z_{F^{\xi^2}} \times \dots \times Z_{F^{\xi^k}}$, $j = 1, \dots, k$, $k = 1, 2, \dots, M$. See that (generally) the multifunctions $\mathcal{K}_{F^{\xi^{k+1}|\bar{\xi}^k}}^{k+1}(\bar{x}^k, \bar{z}^k)$, $k = 1, \dots, M-1$ can depend on the system of the conditional probability measures.

We introduce the M -stage stochastic programming problem ($M \geq 2$) in a rather general form as the problem.

$$\text{Find} \quad \varphi_{\mathcal{F}}(M) = \inf \{ \mathbf{E}_{F^{\xi^1}} g_{\mathcal{F}}^1(x^1, \xi^1) \mid x^1 \in \mathcal{K}^1 \}, \quad (1)$$

where the function $g_{\mathcal{F}}^1(x^1, z^1)$ is defined recursively

$$g_{\mathcal{F}}^k(\bar{x}^k, \bar{z}^k) = \inf\{\mathbb{E}_{F^{\xi^{k+1}}|\bar{\xi}^k=\bar{z}^k} g_{\mathcal{F}}^{k+1}(\bar{x}^{k+1}, \bar{\xi}^{k+1}) | x^{k+1} \in \mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k)\},$$

$$k = 1, \dots, M-1,$$

$$g_{\mathcal{F}}^M(\bar{x}^M, \bar{z}^M) := g_0^M(\bar{x}^M, \bar{z}^M), \quad \mathcal{K}^1 = X^1. \tag{2}$$

$\mathbb{E}_{F^{\xi^1}}, \mathbb{E}_{F^{\xi^{k+1}}|\bar{\xi}^k=\bar{z}^k}, k = 1, \dots, M-1$ denote the operators of mathematical expectation corresponding to $F^{\xi^1}, F^{\xi^{k+1}}|\bar{\xi}^k=\bar{z}^k$.

Evidently, generally, a complete information on the system

$$\mathcal{F} = \{F^{\xi^1}(z^1), \quad F^{\xi^k}|\bar{\xi}^{k-1}(z^k|\bar{z}^{k-1}), k = 2, \dots, M\} \tag{3}$$

is a necessary condition to solve the above defined problem. However in applications very often one of the following cases happen: The system (3) must be replaced by its statistical estimates; the system (3) must be (for numerical difficulties) replaced by some simpler one; the actual system (3) is a little modified (e.g. by a contamination). Consequently, a stability (considered with respect to a probability measures space) and empirical estimates in the multistage problems have been (in the stochastic programming literature) investigated (see e.g. [3], [4], [9]). Approximation solution schemes (in the case of “deterministic” constraints set) have been suggested e.g. in [13] (see also [4]). We focus to a special case when the random element follows an autoregressive random sequence and the constraints sets correspond to a system of the individual probability constraints. In details we assume:

A.1 $\{\xi^k\}_{k=-\infty}^{\infty}$ follows a nonlinear autoregressive sequence

$$\xi^k = H(\xi^{k-1}) + \varepsilon^k, \tag{4}$$

where $\xi^0, \varepsilon^k, k = 1, 2, \dots$ are stochastically independent ($\varepsilon^k, k = 1, \dots$ identically distributed), $H(z)$ is a Lipschitz function on R^s (we denote the distribution function corresponding to ε^1 by the symbol F^ε and suppose the realization ξ^0 to be known),

A.2 there exist $g_i^{k+1}(\bar{x}^{k+1}), \quad k = 1, 2, \dots, M-1$ defined on $R^{n(k+1)}$ and $\alpha_i \in (0, 1), i = 1, \dots, s, \bar{\alpha} = (\alpha_1, \dots, \alpha_s)$ such that

$$\mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k) \quad (:= \mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha})) =$$

$$\bigcap_{i=1}^s \{x^{k+1} \in X^{k+1} : P_{F^{\xi^{k+1}}|\bar{\xi}^k=\bar{z}^k} \{g_i^{k+1}(\bar{x}^{k+1}) \leq \xi_i^{k+1}\} \geq \alpha_i\}. \tag{5}$$

This special case has been already investigated in [5]. In this contribution we go deeper. Our stability results will be based on the $\mathcal{L}_1$ norm and, moreover we intend to investigate (separately) the stability of the constraints sets; the last is important from applications point of view. To this end we employ the results achieved for one-stage problems. To justify this approach see e.g. [3].

2 Some Definitions and Auxiliary Assertions

To achieve stability results for one stage problems very often the Kolmogorov d_K^s and the Wasserstein $d_{W_1}^s$ metrics, $s \geq 1$ (in the probability measures space) have been employed (for the definitions see e.g. [11]). We recall the results based (mostly) on $\mathcal{L}_1$ norm and on an approach of [15]. To this end let $\mathcal{M}_s(R^s) = \{\nu \in \mathcal{P}(R^s) : \int_{R^s} \|z\| \nu(dz) < \infty\}$, $\mathcal{P}(R^s)$ denotes the set of

all Borel probability measures on R^s , $s \geq 1$, $\|\cdot\|$ denotes a suitable norm in R^s ; $\xi := \xi(\omega) = [\xi_1, \dots, \xi_s]$, $\eta := \eta(\omega) = [\eta_1, \dots, \eta_s]$ be s -dimensional random vectors defined on $(\Omega, \mathcal{S}, P)$. We denote by $F, G; P_F, P_G; Z_F, Z_G$ the corresponding distribution functions, probability measures and probability measures supports; by $F_i, G_i, i = 1, \dots, s$ the corresponding one-dimensional marginal distribution functions. Let, moreover $X \subset R^n$ be a nonempty set.

Lemma 1. [7] *Let $P_F, P_G \in \mathcal{M}_1(R^s)$. If for every $x \in X$, g_0 is a Lipschitz (with respect to $\mathcal{L}_1$ norm) function of $z \in R^s$, $z = (z_1, \dots, z_s)$, the Lipschitz constant L is not depending on $x \in X$. If, moreover, for every $x \in X$ a finite $E_F g_0(x, \xi), E_G g_0(x, \eta)$ exist, then*

$$|E_F g_0(x, \xi) - E_G g_0(x, \eta)| \leq L \sum_{i=1}^s \int_{-\infty}^{+\infty} |F_i(z_i) - G_i(z_i)| dz_i \quad \text{for } x \in X.$$

Lemma 2. *Let $X_F, X_G, X \subset R^n$ be nonempty, compact sets, g_0 be a uniformly continuous function defined on $X \times R^s$; $X_F, X_G \subset X$. If, moreover, for every $x \in X$ a finite $E_F g_0(x, \xi), E_G g_0(x, \eta)$ exist, then*

$$\begin{aligned} & |\inf\{E_F g_0(x, \xi) | x \in X_F\} - \inf\{E_G g_0(x, \eta) | x \in X_G\}| \leq \\ & |\inf_{X_F} E_F g_0(x, \xi) - \inf_{X_F} E_G g_0(x, \eta)| + |\inf_{X_F} E_G g_0(x, \eta) - \inf_{X_G} E_G g_0(x, \eta)|. \end{aligned}$$

Proof. The assertion of Lemma 2 follows from the triangular inequality. $\square$

Let f_i , be defined on R^n ; $k_{F_i}(\alpha_i) = \sup_{z_i \in R^1} P_{F_i}\{z_i \leq \xi_i\} \geq \alpha_i$, $k_{G_i}(\alpha_i) = \sup_{z_i \in R^1} P_{G_i}\{z_i \leq \eta_i\} \geq \alpha_i$, $i = 1, \dots, s$. If

$$\begin{aligned} X_F &= X_F(\bar{\alpha}) = \bigcap_{i=1}^s \{x \in X : P_{F_i}\{f_i(x) \leq \xi_i\} \geq \alpha_i\}, \\ X_G &= X_G(\bar{\alpha}) = \bigcap_{i=1}^s \{x \in X : P_{G_i}\{f_i(x) \leq \eta_i\} \geq \alpha_i\}, \end{aligned}$$

then

$$\begin{aligned} X_F &= X_F(\bar{\alpha}) = \bigcap_{i=1}^s \{x \in X : f_i(x) \leq k_{F_i}(\alpha_i)\}, \\ X_G &= X_G(\bar{\alpha}) = \bigcap_{i=1}^s \{x \in X : f_i(x) \leq k_{G_i}(\alpha_i)\}. \end{aligned}$$

Lemma 3. Let P_{F_i} , $i = 1, \dots, s$ be absolutely continuous (with respect to the Lebesgue measure on R^1) probability measures, (we denote by f_{F_i} the corresponding probability densities). If for $i \in \{1, \dots, s\}$

1. there exist constants $\delta_1, \vartheta_F(\bar{\alpha}) > 0$ such that $f_{F_i}(z_i) \geq \vartheta_F(\bar{\alpha})$ for $|z_i - k_{F_i}(\alpha_i)| \leq 2\delta_1$,
2. $|F_i(z_i) - G_i(z_i)| \leq \delta$, $\delta < \delta_1 \vartheta_F(\bar{\alpha})$ for $|z_i - k_{F_i}(\alpha_i)| \leq \delta_1$,

then

$$k_{G_i}(\alpha_i) \in \langle k_{F_i}(\alpha_i) - \frac{\delta}{\vartheta_F(\bar{\alpha})}, k_{F_i}(\alpha_i) + \frac{\delta}{\vartheta_F(\bar{\alpha})} \rangle,$$

$$X_G(\bar{\alpha} + \bar{\delta}) \subset X_F(\bar{\alpha}) \subset X_G(\bar{\alpha} - \bar{\delta}), \quad \bar{\delta} = (\delta, \dots, \delta).$$

Proof. The proof of the assertions follows immediately from the assumptions and from the definitions of $k_{F_i}(\alpha_i)$, $k_{G_i}(\alpha_i)$, $i = 1, \dots, s$. $\square$

Lemma 4. Let $\xi^1, \xi^2, \varepsilon^2$ follow the assumption A.1. If

1. P_{F^ε} is absolutely continuous with respect to the Lebesgue measure on R^s ,
2. $h(z^2, z^1)$ is a Lipschitz function on $Z_{F^{\varepsilon^2}} \times Z_{F^{\varepsilon^1}}$ (w.r.t. $\mathcal{L}_2$ norm); there exists a finite $E_{F^{\varepsilon^2}|\xi^1} h(\xi^2, \xi^1)$,

then $E_{F^{\varepsilon^2}|\xi^1=z^1} h(\xi^2, \xi^1)$ is a Lipschitz function on $Z_{F^{\varepsilon^1}}$ (w.r.t. $\mathcal{L}_2$ norm).

Proof. The proof of Lemma 4 follows from the elementary properties of the integral. $\square$

Lemma 5. [2] Let $Y \subset R^m$, $X \subset R^n$ be nonempty sets. If

1. $\bar{\mathcal{K}}(y)$ is a multifunction mapping Y into the space of nonempty, closed subsets of X such that it holds for a constant $\hat{C}$

$$\Delta[\bar{\mathcal{K}}(y(1)), \bar{\mathcal{K}}(y(2))] \leq \hat{C} \|y(1) - y(2)\| \quad \text{for every } y(1), y(2) \in Y,$$

2. $\bar{h}(x, y)$ is a Lipschitz function on $X \times Y$ with the Lipschitz constant L ,
3. for every $y \in Y$, $\inf\{\bar{h}(x, y) : x \in \bar{\mathcal{K}}(y)\} > -\infty$,

then

$$\varphi(y) = \inf\{\bar{h}(x, y) : x \in \bar{\mathcal{K}}(y)\}$$

is a real-valued Lipschitz function on Y with the Lipschitz constant $L(\hat{C}+1)$. ($\Delta[\cdot, \cdot]$ denotes the Hausdorff distance, for the definition see e.g. [12].)

3 Stability and Approximation Analysis

If F^ε is replaced by another s -dimensional distribution function G^ε , then we obtain instead of the system $\mathcal{F}$ another system $\mathcal{G}$

$$\mathcal{G} = \{G^{\xi^1}(z^1), G^{\xi^k|\xi^{k-1}}(z^k|\bar{z}^{k-1}), k = 2, \dots, M\} \quad (6)$$

and a new multistage stochastic programming problem. We denote its optimal value by $\varphi_{\mathcal{G}}(M)$. Applying the results of the previous section and the results of [7], we can obtain assumptions under which there exist constants C_{W_1} , C_K , $C_{W_1}^i$, C_K^i , $i = 1, \dots$ such that

$$\begin{aligned} |\varphi_{\mathcal{F}}(M) - \varphi_{\mathcal{G}}(M)| &\leq C_{W_1} d_{W_1}^s(P_{F^\varepsilon}, P_{G^\varepsilon}) + C_K \max_{i \in \{1, \dots, s\}} d_K[P_{F_i^\varepsilon}, P_{G_i^\varepsilon}], \\ |\varphi_{\mathcal{F}}(M) - \varphi_{\mathcal{G}}(M)| &\leq \sum_{i=1}^s C_{W_1}^i \int_{R^1} |F_i^\varepsilon(z_i) - G_i^\varepsilon(z_i)| dz_i + \sum_{i=1}^s C_K^i |k_{F_i^\varepsilon} - k_{G_i^\varepsilon}|. \end{aligned} \quad (7)$$

The first inequality in (7) (with the underlying $\mathcal{L}_2$ norm) follows from [3] or [5]. A similar assertion can be proven when $\{\xi^k\}_{k=-\infty}^\infty$ follows a Markov sequence. However, stronger assumptions on the probability densities has been then assumed (for details see e.g. [3]). The second inequality employs $\mathcal{L}_1$ norm in R^s . The idea to employ $\mathcal{L}_1$ norm appears first in [10] and furthermore in [13]. However, the case of a discrete approximation of the system $\mathcal{F}$ and the constraints sets given by a system of algebraic inequalities has been there considered. The results corresponding to Lemma 1 for general G has been proven in [7]. This new result gives, furthermore, possibility to apply the properties of integrated empirical process to the empirical estimates in stochastic programming problems (for details see e.g. [6], [7], [14]).

Evidently, employing the second inequality, we can construct solution approximation schemes on the basis of discretization the “underlying” one-dimensional marginal distribution functions F_i^ε , $i = 1, \dots, s$. The constants $C_{W_1}^i$, d_K^i , $i = 1, \dots, s$ can be estimated on the basis of the former results (see e.g. [2]). However to construct responsible solution approximation schemes it is necessary to estimate the value δ such that for $k = 1, \dots, M - 1$,

$$\mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha} + \bar{\delta}) \subset \mathcal{K}_{\mathcal{G}}^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha}) \subset \mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha} - \bar{\delta}).$$

To this end we define for $k = 1, \dots, M - 1$, $i = 1, \dots, s$

$$\begin{aligned} k_{F_i^{\varepsilon^{k+1}|\xi^k=z^k}}(\alpha_i) &= \sup \{z_i^{k+1} \in R^1 : P_{F_i^{\varepsilon^{k+1}|\bar{\xi}^k=\bar{z}^k}}\{z_i^{k+1} \leq \xi_i^{k+1}\} \geq \alpha_i\}, \\ k_{F_i^\varepsilon}(\alpha_i) &= \sup \{z_i \in R^1 : P_{F_i^\varepsilon}\{z_i \leq \varepsilon_i\} \geq \alpha_i\}. \end{aligned}$$

Proposition 1. *Let F^ε , G^ε be two s -dimensional distribution functions. Let, moreover, $P_{F_i^\varepsilon}$, $i = 1, \dots, s$ be absolutely continuous (with respect to the Lebesgue measure on R^1) probability measures. We denote by $f_{F_i^\varepsilon}$ the corresponding probability densities, $i = 1, \dots, s$. If*

1. *there exist constants δ_1 , $\vartheta_{F^\varepsilon}(\bar{\alpha}) > 0$ such that $f_{F_i^\varepsilon}(z_i) \geq \vartheta_{F^\varepsilon}(\bar{\alpha})$ for $|z_i - k_{F_i^\varepsilon}(\alpha_i)| \leq 2\delta_1$, $i = 1, \dots, s$,*
2. *$|F_i^\varepsilon(z_i) - G_i^\varepsilon(z_i)| \leq \delta$, $\delta < \delta_1 \vartheta_{F^\varepsilon}(\bar{\alpha})$ for $|z_i - k_{F_i^\varepsilon}(\alpha_i)| \leq \delta_1$,*

then for $\bar{x} \in \bar{X}^k$, $\bar{z}^k \in \bar{Z}_{\mathcal{F}}^k$, $k = 1, \dots, M$,

$$k_{G_i^\varepsilon}(\alpha_i) \in \langle k_{F_i^\varepsilon}(\alpha_i) - \frac{\delta}{\vartheta_{F(\bar{\alpha})}}, k_{F_i^\varepsilon}(\alpha_i) + \frac{\delta}{\vartheta_{F(\bar{\alpha})}} \rangle, \quad i = 1, \dots, s,$$

$$\mathcal{K}_G^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha} + \delta) \subset \mathcal{K}_{\mathcal{F}}^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha}) \subset \mathcal{K}_G^{k+1}(\bar{x}^k, \bar{z}^k; \bar{\alpha} - \delta).$$

Proof. Since $k_{F_i^\varepsilon}(\alpha_i) = k_{F_i^{\varepsilon^k | \bar{\varepsilon}^k - 1 = \bar{z}^k - 1}}(\alpha_i) - H_i(z^{k-1})$, $\bar{z}^k \in R^{s^k}$, $i = 1, \dots, s$, $k = 2, \dots, M$, the assertion follows from the assertion of Lemma 3. $\square$

Acknowledgement. The research was supported by the Grant Agency of the Czech Republic under Grants 402/04/1294, 402/05/0115 and 402/06/1417.

References

1. Dupacová J (1995) Multistage stochastic programs: the state-of-the-art and selected bibliography. *Kybernetika* 31: 151–174.
2. Kanková V (1998) A note on multistage stochastic programming. In: *Proceedings of 11th joint Czech–Germany–Slovak Conf.: Mathematical Methods in Economy and Industry*. University of Technology, Liberec (Czech Republic) 1998, pp 45–52.
3. Kanková V (2002) A Remark on the analysis of multistage stochastic programs, Markov dependence. *Z angew Math Mech* 82:781–793.
4. Kanková V, Smíd M (2004) On approximation in multistage stochastic programs: Markov dependence. *Kybernetika* 40:625–638.
5. Kanková V (2006) Decomposition in multistage stochastic programming programs with individual probability constraints. In: *Operations Research Proceedings 2005*. Springer, Berlin: 793–798.
6. Kanková V, Houda M (2006) Dependent samples in empirical estimation of stochastic programming problems. *Austrian Journal of Statistics* 35:271–279.
7. Kanková V, Houda M (2006) Empirical estimates in stochastic programming. In: *Proceedings of Prague Stochastics 2006* (M. Husková and M. Janžra, eds.). MATFYZPRESS, Prague: 426–436.
8. Kuhn D (2005) Generalized bound for convex multistage stochastic programs. *Lectures Notes in Economics and Mathematical Systems*. Springer, Berlin.
9. Mänz A, Vogel S (2006) On stability of multistage stochastic decision problem. In: *Recent Advances in Optimization. Lecture Notes in Economics and Mathematical Systems* (A. Seeger, ed.). Springer, Berlin.
10. Pflug GCh (2001) Scenario tree generation for multiperiod financial optimization by optimal discretization. *Math Program ser B* 89: 251–271.
11. Rachev ST(1991) *Probability metrics and the stability of stochastic models*. Wiley, Chichester.
12. Salinetti G (1983) Approximations for chance constrained programming problems. *Stochastics* 10:157–179.
13. Smíd M (2005) On Approximation of stochastic programming problems. Doctoral Thesis, Charles University, Prague.
14. Shorack GR, Wellner JA (1986) *Empirical processes with applications to statistics*. Wiley New York.
15. Valander SS (1973) Calculation of the Wasserstein distance between probability distributions on the line (in Russian). *Theor Probab Appl* 18: 784–786.

ALM Modeling for Dutch Pension Funds in an Era of Pension Reform

Willem K. Klein Haneveld, Matthijs H. Streutker, and
Maarten H. van der Vlerk

Dept. of Econometrics & OR, University of Groningen, The Netherlands
`m.h.streutker@rug.nl`

Summary. This paper describes the features of a multistage ALM recourse model in light of the recent pension reform in the Netherlands. The main results are explicit modeling of indexation decisions and modeling of new regulatory rules.

1 Introduction

Pension fund management in the Netherlands is on the move, which sets ALM modeling for pension funds for new challenges. Years of apparent stability abruptly ended with the severe drop of the stock indices at the beginning of the new millennium. This unrest, also fed by the ageing discussion, encouraged a broadly conducted debate on the Dutch pension system. Important outcomes of the debate are trimmed-down pension contracts and the development of new regulatory rules for pension funds. Currently, most pension contracts are based on average earnings instead of final earnings as in the past, and full indexation (correction for inflation) is not granted anymore. The new regulatory rules, which are called the Financieel Toetsingskader (Financial Assessment Framework) and are scheduled to be implemented in January 2007, are based on risk based supervision and show strong affinity with the Solvency [6] and Basel [2] Accords.

An upcoming ALM methodology is multistage stochastic programming. The headway made is described in surveys by Mulvey [9] and Sodhi [11]. In this paper we show how we deal with the recent pension reform in our multistage recourse ALM model.

2 Pension Reform

The pension reform takes place on two fronts. First, in the past ten years the pension contracts have been extremely trimmed-down by (1) the shift from

final to average earnings contracts and (2) absence of complete indexation. Second, new far-reaching regulatory rules for pension funds are bound to be implemented.

2.1 Trimmed-Down Pension Contracts

Almost all pension contracts in the Netherlands are defined benefit contracts, where the benefit payments upon retirement are predetermined, as they are linked to the salary of the participant. Each working year the participant accrues new rights, for which contributions are paid by the participant and his employer. In the past in the majority of the contracts the benefit payments were based on the final salary before retirement. Such final earnings (FE) schemes lead to a pension that is 70% of the final salary. If this 70% goal is to be reached in 40 years, then the so-called accrual rate is $\frac{70\%}{40} = 1.75\%$. In the past years the pension schemes changed to the more sparing average earnings (AE) schemes. In these schemes the benefit payments are based on the average salary instead of final salary. In Table 1 the percentage of active participants per scheme are given for the last 5 years. In 2001 the majority of the contracts were FE contracts. However, in only 5 years time this completely shifted to AE schemes.

Table 1. Percentage of active participants in the Netherlands with a final earnings (FE) and an average earnings (AE) scheme. Source: DNB [3]

Year	FE	AE
2001	56.1	32.2
2002	54.3	31.6
2003	49.3	35.4
2004	12.0	72.6
2005	10.6	74.3

Another important development is that the common practice of granting full indexation stopped. The nominal rights as accrued are guaranteed, but the correction of the benefit rights for inflation is conditional. The indexation of ABP[1], which is one of the largest pension funds in the Netherlands, for the past two years did not equal the wage inflation. It was 79% (2004) and 45% (2005) of the wage inflation. As there is in general a long period between accrual and consumption of rights, the indexation of rights is essential for a good pension, especially in an AE scheme as pointed out by Van Ewisk [5].

To underline the effects of these two developments, we illustrate the typical build-up of pension rights in a FE and a (fully-indexed) AE scheme in Fig. 1. The parameters are a starting salary of 10,000 Euro, every 5 years a 10% salary jump due to career improvements, 2.4% wage inflation each year, and an accrual rate of 1.75%. The FE scheme results in the 70% of final salary. The

pension under the fully-indexed AE scheme is only 51% of the final salary. When no indexation is granted the pension is only 36% of the final salary, which is about half of the FE scheme.

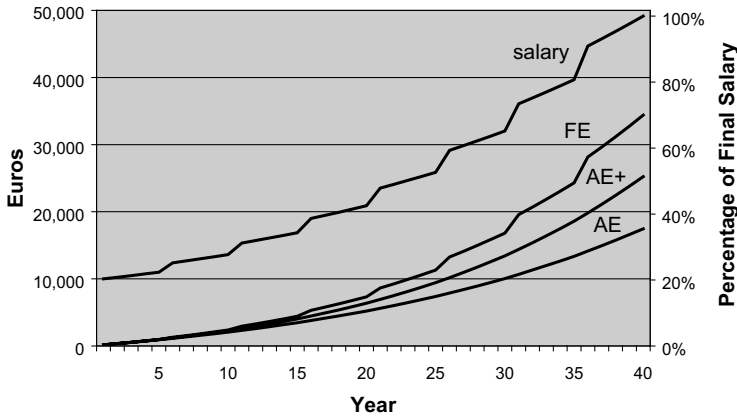


Fig. 1. Build-up of rights in a final earnings (FE), fully-indexed average earnings (AE+), and nominal average earnings (AE) scheme

2.2 New Regulatory Rules

The Financieel Toetsingskader (FTK) [10] states that it is intended to improve the insight of both the supervised institution as well as the supervisory authority into the institution's financial position and its possible development over the short and medium term. Its main goal is to make the ins and outs of pension funds more transparent. The FTK is involved with the subjects of valuation of liabilities, solvency of the fund, contribution rate, and indexation. Below, we will outline these four fields, for more information see [7].

First, for the valuation of the liabilities the FTK promotes the use of an interest rate term structure. More specifically, the FTK prescribes the use of an interest rate swap curve. Second, the FTK has two conditions on the solvency (assets minus liabilities). First, the solvency should be at least 5% of the liabilities. If not, a recovery period of one year is allowed. Second, the solvency should be such that the probability of underfunding in the next year is at most 2.5%. If not, a recovery period of 15 years is allowed. Third, the contribution rate must be cost-effective, and moreover, the FTK would like the contribution rate to be stable in order to prevent social upheaval. Fourth, the indexation policy must be in line with the indexation ambition communicated to the participants. Depending on this ambition, funds must construct buffers and can ask contribution for indexation. Further, in case of financial distress funds must be cautious with indexation.

3 ALM Modeling Implications

In this section we show how we deal with the above mentioned developments in our multistage recourse model. First, we will explain our approach on indexation. Second, our modeling of the FTK is outlined.

3.1 Indexation

In almost all existing Dutch ALM models the liabilities are seen as data, i.e., there are no decisions on the liability side. In the past this was defensible, since (1) full-indexation was common practice, and (2) the final earnings scheme was the dominant pension contract and in those contracts incomplete indexation only affects the passive participants. However, given the current practice of average earnings schemes and incomplete indexation, the modeling of indexation decisions is necessary. Moreover, the FTK has increased the attention on indexation even more. Drijver [4] was the first to occupy indexation in his model. However, he uses a rather rough approximation approach.

As described in [8] we model the indexation process accurately by relating the value of the nominal liabilities to the period the rights were accrued. Introducing parameters

$\underline{L}_t^s$ = value of nominal liabilities originating from rights accrued in period s , and decision variables

$\underline{L}_t^s$ = accumulated indexation of rights accrued in period s ,

we get the following value of the liabilities at time t

$$L_t = \sum_{s=s_t}^t I_t^s \underline{L}_t^s$$

with the oldest rights at time t stemming from period s_t . The accumulated indexation decisions are restricted in the following way:

$$I_t^s \leq I_t^{s-1} \leq w_s I_t^s \tag{1}$$

$$I_t^{s-1} - I_t^s \geq I_{t-1}^{s-1} - I_{t-1}^s \tag{2}$$

with w_s the wage increase factor for period s . Constraint (1) models that the nominal rights are guaranteed and that the indexation in a certain period cannot be higher than the wage inflation in that period. Constraint (2) is a linear relaxation of the restriction that once granted indexation cannot be taken back.

Now that the liabilities are decisions, they need to be directed in the objective. We choose to model the subobjectives of the stakeholders involved (active participants, passive participants, employer) separately. In order to be able to combine them to one, we formulate the interests of a stakeholder group as the discounted expected cash flows from/to this group. Here we will go

into the subobjective of the passive participants (retirees, deferred members). These participants do not pay contributions anymore, so their interests are the expected benefit payments they receive, which are influenced by the indexation decisions. Let $\underline{B}_t^s(u)$ be the expected nominal benefit payments at time t to passive participants in period u originating from rights accrued in period s , then the discounted value of these expected payments at $t = 0$, i.e., the beginning of the decision period, is given by

$$\underline{L}_{t,0}^s = \sum_{u=t+1}^U d_o(u) \underline{B}_t^s(u)$$

with $d_o(u)$ the discount factor at time 0 for payments in period u and U the last period of expected benefit payments. The actual value of the expected benefit payments and thus the value of the liabilities depends on the indexation decision. The gain in expected cash flows at a decision time t is given by

$$\sum_{s=s_t}^{t-1} \left(I_t^s - I_{t-1}^s \right) \underline{L}_{t,0}^s.$$

The subobjective of the passive participants is now given by

$$\max_{I_t^s, \dots} \sum_{t=1}^T \sum_{s=s_t}^t \left(I_t^s - I_{t-1}^s \right) \underline{L}_{t,0}^s$$

with T the horizon of the decision model.

3.2 FTK Implications

Since we aim at applicability of our model in practice, we strive to accurately model the requirements as set out in the FTK. Moreover, we would like to contribute to the discussions on the FTK. Many believe that the FTK is too strict and might result in an eventually abandoning of the defined benefit system in favor of a defined contribution system. We believe that we can give valuable quantitative arguments, since our model will allow to analyze the effects of constraints reflecting the FTK.

The adaptation of the interest rate swap curve for the valuation of the liabilities as prescribed by the FTK is too complex for a multistage recourse model. We resort to a government yield curve, which satisfies, like the swap curve method, valuation of liabilities based on a risk-free financial product.

The FTK has a short term as well as a medium term solvency requirement. In [8] we explained that we model the short term requirement using binary variables and that we are considering various alternatives, including chance constraints and integrated chance constraints, for the medium term solvency requirement.

The need of the FTK for a cost-effective contribution rate is met, since our model is an optimization model. Now that the liabilities are valued using an interest rate term structure, the volatility of the solvency is increased, which could incur volatile contribution rates. In order to preserve a stable contribution rate, we bound the contribution rate and changes in the rate.

The regulatory rules on indexation have not been crystallized out yet. The supervisor and pension funds are searching for accurate formulations of indexation ambitions. For now we do not include any constraints on indexation besides the ones mentioned in the previous section.

4 Conclusion

We described the recent pension reform in the Netherlands and its modeling implications for multistage recourse ALM models. We showed how we model the important indexation decisions and outlined our adoption of the new regulatory rules.

References

1. Algemeen Burgerlijk Pensioenfonds (ABP). <http://www.abp.nl>.
2. Basel Committee on Banking Supervision. Basel accords. <http://www.bis.org/bcbs/index.htm>.
3. De Nederlandsche Bank (DNB). <http://www.dnb.nl>.
4. S.J. Drijver. *Asset Liability Management for Pension Funds Using Multistage Mixed-Integer Stochastic Programming*. PhD thesis, University of Groningen, 2005.
5. C.D. van Ewijk. Reform of occupational pensions in the Netherlands. *De Economist*, 153(3):331–347, 2005.
6. European Insurance and Occupational Pensions Committee. Solvency accords. http://ec.europa.eu/internal_market/insurance/solvency_en.htm.
7. W.K. Klein Haneveld, M.H. Streutker, and M.H. van der Vlerk. Implementation of new regulatory rules in a multistage ALM model for Dutch pension funds. SOM research report, University of Groningen, forthcoming.
8. W.K. Klein Haneveld, M.H. Streutker, and M.H. van der Vlerk. Modeling ALM for Dutch pension funds. In J. Safrankova, editor, *WDS'06 Proceedings of Contributed Papers: Part I - Mathematics and Computer Sciences*, Matfyzpress, Prague, forthcoming.
9. J.M. Mulvey. Introduction to financial optimization: Mathematical programming special issue. *Mathematical Programming*, 89:205–216, 2001.
10. PVK. Financial assessment framework consultation document. http://www.dnb.nl/dnb/bin/doc/FTK%20Consultation%20Document%20English%20translation_tcm13-47968.pdf, October 2004.
11. M.S. Sodhi. LP modeling for asset-liability management: A survey of choices and simplifications. *Operations Research*, 53(2):181–196, 2005.

System Dynamics and Dynamic Modelling

Identifying Fruitful Combinations Between System Dynamics and Soft OR

Myrjam Stotz and Andreas Größler

Industrieseminar, Mannheim University, Germany

MyrjamStotz@gmx.de

agro@is.bwl.uni-mannheim.de

Summary. Since the 1960s qualitative problem structuring methods have evolved mainly in the UK from traditional operational research, which are called „soft OR“. These methods focus on generating understanding of systems as a whole, on improving systems, and on personal learning. System dynamics has similar goals; therefore, it might be useful to combine it with soft OR. Some examples of combinations exist; however, these are mainly anecdotal. Thus, systematic research promises new insights and knowledge about the useful combination of soft OR and system dynamics in practice and theory. This has often been asked for but has not yet been established. In this paper a characterization of system dynamics is presented that can be used to find potential combinations of system dynamics and soft OR. More precisely, a framework designed by [11] for mapping methods is described. This framework characterizes methods along two dimensions: four steps of intervention processes and three domains, in which problem situations manifest: social, personal, and material world. System dynamics is mapped and thus characterized by this framework. On the one hand, the mapping highlights the strengths of system dynamics; on the other hand, it shows possibilities for useful combinations with soft OR. With this study, further steps to a systematic research of combinations between system dynamics and soft OR are undertaken. Based on the method characterization of system dynamics and soft OR, proposals for concrete combinations can be derived and contextual factors for successful combinations can be investigated.

1 A Framework for Mapping System Dynamics and Soft OR

The term „soft OR“ describes a variety of methods, which help to structure complex and problematic situations [16]. Soft OR accepts subjective views of participants and deals with uncertainty and conflict between the stakeholders. It tries to resolve conflicting viewpoints among stakeholders and helps to achieve consensus about future action [16], [17]. However, with the focus on qualitative modeling, soft OR lacks quantitative techniques like simulation to

verify the subjective assumptions made in qualitative group discussions and to examine the time-dependent behavior of systems [3], [8].

In contrast to soft OR, it is one of the core elements of system dynamics to build a quantified simulation model from the description of the problem situation. Simulation runs with a model lead to understanding and improvement of the dynamic behavior of the system [21]. But, as [3, p.3] states, "little guidance exists for converting a real-life situation into a simulation model". With its focus on problem structuring, this is a strength of soft OR.

[8] as well as [13] describe combinations of soft OR and system dynamics and discuss the potential of the combination. They show that not only in modeling itself but also in other stages of a problem solving process, e.g. problem conceptualization or scenario analysis, the combined use of system dynamics and soft OR makes sense. As examples [8] mentions: strategic options development and analysis (SODA) giving rise to a system dynamics model [2], soft systems methodology (SSM) to negotiate the purpose of a system [9] or robustness analysis used to explore the scenarios and decision analysis to evaluate their desirability [4], [15].

Mostly, system dynamics is combined with SSM or cognitive mapping [13]. However, systematic research of possible combinations has often been asked for [8], [3], [11], [14], but has not yet been undertaken.

[11] proposes a framework for mapping methods with the aim of identifying suitable method combinations. The framework has two dimensions. The first one consists of four different stages of an intervention process: (1) appreciation of the problem situation, (2) analysis of the underlying structure and constraints generating the situation, (3) assessment of alternatives and finally, (4) action to bring about desirable changes. Of course, these stages are not seen as discrete sequential steps, but they represent different aspects that need to be considered throughout the process.

The second dimension represents the multi-dimensionality of problem situations. The fact that each problem has different aspects is represented by three "worlds": the material, personal, and social world. With this distinction Mingers refers to Habermas [5, 6]. The material world is the world which exists independently from human beings. The personal world is the world of the thoughts, emotions, experiences and beliefs of individuals. The social world is the world that individuals as members of particular social systems share. It results from interactions and relationships between human beings. Following this definition, the term social world represents in a general way social structures between individuals and is not limited, for instance, to activities in the social or health care sector. In the same way, institutions (like hospitals or schools) per se do not belong to the social world as defined by Mingers, they rather belong to the material world.

The framework can be used in two ways: On the one hand, relative strengths of a method compared with others can be highlighted. On the other hand, requirements on selecting methods for a concrete problem situation can be defined. As each method has different strengths and each problem has dif-

ferent requirements, methods that fit best with the problem situation under consideration can be chosen. Thus, the framework helps to choose appropriate methods and combinations in a specific problem situation. Each stage of the project can be supported by suitable methods [11].

2 Mapping System Dynamics

Mingers’ framework described above is a useful way to find suitable combinations of soft OR and system dynamics. Therefore, system dynamics is mapped by the framework to show fields in which system dynamics and soft OR can be combined to supplement each other. [10] shortly characterizes systems dynamics as a method, which is particular strong in supporting problem situations concerning the material world. But it has more potentials as shown in the following section. Of course, the mapping of system dynamics is debatable and only refers to the ”normal” usage of system dynamics. But this mapping is not about evaluating and ranking methods. Its purpose is only to lead to potential combinations of methods, which complement or support each other. The map for system dynamics is shown in Figure 1. The darker the fields are shaded, the better system dynamics fits for the described tasks.

	(1) Appreciation of	(2) Analysis of	(3) Assessment of	(4) Action to
Social World	roles, norms, social practices, culture and power relations	underlying social structures	ways of changing existing practices and culture	generate enlightenment of social situation and empowerment
Personal World	individual beliefs, meanings, values, and emotions	differing Weltanschauungen ¹ and personal rationalities	alternative conceptualizations and constructions	generate understanding, personal learning and accommodation of views
Material World	material and physical processes and arrangements	underlying causal structures	alternative physical and structural arrangements	select and implement best alternatives

Fig. 1. System dynamics in Mingers’ framework

(1) Appreciation of the problem situation

In a system dynamics project, the focus lies on describing the system and building a model that generates problematic system behavior. To appreciate the problem situation, the problem has to be defined clearly from a dynamic perspective, key variables have to be found, and structures between these key variables have to be figured out. In this phase often causal loop diagrams are built that represent the underlying (feedback) structure which generates the

problematic behavior of a system. In the next stages of a system dynamics project, a quantitative model is built. Variables are represented as stocks and flows, i.e. levels and rates. In this way, the accumulations and changes of material, money and information are symbolized, which makes system dynamics strong in appreciating material and physical processes. It thus appreciates the material world.

In the context of the social world, system dynamics is not so strong. Usually, system dynamics is not used to uncover social structures and to change for example power relations. There are no process stages or techniques, in which the focus lies on finding out which roles, norms, social practices, culture, and power relations exist. Social structures are only indirectly appreciated when they influence human behavior that has an effect on the model variables. Such effects are represented in system dynamics models as policies which influence the physical structures [21]. An example for this is Senge's model of the U.S.-U.S.S.R. arms race[19]. It deals with the power relations between the USA and the Soviet Union during the Cold War.

The appreciation of individual beliefs and emotions are also not in the focus of a system dynamics study. Of course, causal loop diagrams represent mental models of individuals. But in system dynamics projects it is usually worked with only one diagram which represents a shared mental model of all participants [7]; it is not the aim to generate multiple models and to find out which and how many different viewpoints exist.

(2) Analysis of the underlying structure and constraints generating the problem situation

Following the appreciation of the problem, the structure and constraints generating the problem situation are analyzed. As explained before, originally, the focus of system dynamics is not on appreciating social structures. Hence social structures are normally not analyzed to find a solution to a certain problem situation. Here again, they are only interesting when they influence the physical structure.

Analysis of the personal world in this framework means analyzing differing Weltanschauungen and personal realities. It means exploring in detail the different assumptions and views individuals have of a certain problem situation. This could be compared to building for each viewpoint one separate system dynamics model. However, only one system dynamics model is usually built and simulated. Afterwards changes in the model structure are made which change the model behavior. But it is still only one model that is adapted during the progress of the system dynamics project.

As stated before, from a first qualitative analysis of relevant variables and structures between them, quantitative models are built, level rate equations are formulated that track accumulations of material, money and information as they move through a system. This is one of the outstanding qualities of system dynamics. Ideas, discussions and analysis of causal relationships do not

stay on a qualitative level. Next to qualitative systems thinking, quantitative analysis is an essential part of system dynamics. It is the only practical way to test qualitative models, which can always be colored by unconscious bias [21]. Therefore system dynamics is strong in analyzing causal structures.

(3) Assessment of alternatives

As described above, changes in system structures are tested in scenarios. Simulations visualize how the system behaves with different model structures or values of parameters. Using computer simulations is much quicker than analyzing reactions on changes of the system in real time. Thus, learning processes can be accelerated. Unsuccessful options can be sorted out before implementing them in reality. Ways can be found how the situation can be changed. Thus it is again a particular strength of system dynamics to assess alternative physical and structural arrangements of the material world.

In simulations not only values of parameters and structures are changed, but also new strategies are created and a wide range of alternative scenarios can be analyzed [21]. At this point, alternative conceptualizations and different ideas of individuals can be tested and assessed. They are the input for the scenarios. Thus, scenarios do not only lead to an assessment of possible changes in the material or physical structure, but also to an assessment of different conceptualizations of individuals. However, system dynamics does not assess personal rationalities as a whole (including building system dynamics models and simulating each viewpoint separately); only individual ideas and assumptions of how the system could be improved and how it reacts on changes can be assessed by scenarios.

Maybe it seems to be puzzling that the appreciation and analysis of the personal world is not seen as a strength of system dynamics. It is not the intention of system dynamics to find out all different viewpoints that could exist in a particular problem situation and to analyze how exactly all the individuals think. But, it is intended in system dynamics projects to learn and to test different assumptions the participants have of how a system reacts on changes. An example for this is the use of system dynamics based simulators [7], [18].

As it is not the aim to appreciate and to analyze social structures in most system dynamics projects, regularly, it is also not intended to find ways to change existing social structures or cultures.

(4) Action to bring about desirable changes

Action to bring about desirable changes in the social world in this framework means generating enlightenment of the social situation and empowerment. As system dynamics is a simulation method that aims at improvement of systems and learning, it is not a methodology that has the capabilities to change social structures. Maybe discussions and also simulations could lead to think about social structures (as for example described in Senge's example mentioned above) and could generate common sense to change existing

structures. But system dynamics would not directly support actions to bring about changes of social structures.

System dynamics is also not a methodology for directly implementing changes concerning the material world. It can generate understanding and help convincing others of the necessity of the changes [20]. But there is no technique that supports the required implementation of structural changes. System dynamics can also not fulfill the requirement "selecting best alternatives" proposed by the framework. With simulations, alternative structures can be tested and assessed. However, there exist no rules and techniques, how 'best' alternatives can be selected and implemented. System dynamics can be used to select the best alternative from those tested. But as there is no algorithm in testing all possible alternatives and calculating the optimum, system dynamics is not suitable and not intended to find the 'optimal', the 'best', alternative [3]. However, system dynamics is, of course, useful to assess alternatives by simulations and thus to identify desirable changes.

It is already stated, that different assumptions of individuals can be assessed in simulations. Scenarios quantify and kind of objectify the qualitative assumptions made by individuals. It is tested, whether the expected reactions result of different changes in the model structure or not. Often, intuitively expected behavior is not confirmed by the simulations. As the complexity of the mental models often exceeds the capacity to understand their implications, as they are too large to simulate them mentally, the strength of system dynamics that lies in simulation can be used to generate personal learning, understanding, and the accommodation of views. Different assumptions of individuals can be tested and either be verified or rejected [7], [21]. Thus, even conflicting viewpoints can be aligned. This is again a particular strength of system dynamics.

This characterization of system dynamics shows, that it is especially strong in appreciating, analyzing and assessing aspects of the material world. It is also very well suited to generate understanding, personal learning and accommodate views and thus in action to the personal world. To an extent it can also be used to assess the personal world and for acting to the material world in the sense of selecting "best" alternatives. However, especially to appreciate and analyze the personal world, as well as in projects where the focus lies on the social world, combinations with other methods make sense.

3 A Combination Between System Dynamics and Soft OR

With the mapping of system dynamics it gets clear that system dynamics does not fit best with every problem situation. In these cases, ideas for useful combinations can be generated by the framework.

Cognitive Mapping, a method of soft OR, is for example mapped by [12] as shown in Figure 2. It has strengths in appreciating and analyzing individual

patterns of belief, and in gaining commitment to action, but is relative weak in assessing possible alternatives.

	(1) Appreciation of	(2) Analysis of	(3) Assessment of	(4) Action to
Social World	roles, norms, social practices, culture and power relations	underlying social structures	ways of changing existing practices and culture	generate enlightenment of social situation and empowerment
Personal World	individual beliefs, meanings, values, and emotions	differing Weltanschauungen and personal rationalities	alternative conceptualizations and constructions	generate understanding, personal learning and accommodation of views
Material World	material and physical processes and arrangements	underlying causal structures	alternative physical and structural arrangements	select and implement best alternatives

Fig. 2. Cognitive mapping in Mingers’ framework

A comparison between the map of system dynamics Figure 1 and that of cognitive mapping Figure 2 shows, that both methods could complement each other. It implies that, depending on the problem situation, a combination between system dynamics and cognitive mapping can make sense. One example of a successful combination of cognitive mapping and system dynamics is described by [1]. In their project, cognitive mapping and system dynamics were used to support a litigation claim dealing with delays and disruptions in a very large project. Cognitive mapping was used to elicit and to document the knowledge from involved individuals. It was used to bring together all plausible explanations for the disruptions in the project. Later it was used to find those problem areas that were likely to be the most important causes of the problem. System dynamics was then used to model these problem areas and to illustrate by simulations, how the delays under considerations were caused and how they could have been avoided. In this case, the combination of system dynamics and soft OR proved to be very useful. The strengths of both methods were used. With this work, further steps to a systematic research of combinations between system dynamics and soft OR have been undertaken. System dynamics is mapped according to three dimensions of problem situations and different stages of an intervention. Of course, this can be discussed and new insights may result in a different mapping. But the mapping provides a basis for finding methods that can be combined with system dynamics in a way that the methods complement and support each other. As an example, system dynamics and cognitive mapping are both discussed in the context of Mingers’ framework. A comparison based on this framework shows that a combination of system dynamics and cognitive mapping is useful.

References

1. Ackermann F, Eden C, Williams, T. (1997) Modeling for Litigation Interfaces 27(2):48–65.
2. Eden C. (1989) Using Cognitive Mapping for Strategic Options Development and Analysis (SODA). Rational Analysis for a Problematic World, ed. J. Rosenhead.
3. Forrester J.W (1994) System Dynamics, System Thinking, and Soft OR. System Dynamics Review 10(2):245–256.
4. Gardiner P.C, Ford A (1980) Which Policy Run Is Best?, And Who Says So? TIMS Studies in the Management Science 14: 241–257.
5. Habermans J (1984) Communicative Action Vol. 1: Reason and the Rationalization of Society. London.
6. Habermans J (1987) The Theory of Communicative Action Vol. 2: Lifeworld and System: A Critique of Functionalist Reason. Oxford.
7. Kim D.H (1993) The Link between Individual and Organizational Learning. Sloan Management Review 35(1):37–15.
8. Lane D.C (1994) With a Little Help from Our Friends: How System Dynamics and Soft OR Can Learn from Each Other. System Dynamics Review 10(2-3):101–134.
9. Lane D.C, Oliva R (1994) The Greater Whole: Towards a synthesis of SD and SSM. Proceedings of the 1994 International System Dynamics Society
10. Mingers J (2000) Variety is the spice of life: Combining Soft and Hard OR/MS Methods International Transactions in Operational Research 7:673–691.
11. Mingers J (2001) Multimethodology - Mixing and Matching Methods. Rational Analysis for a Problematic World Revisited. In: Rosenhead J, Mingers J
12. Mingers J(1997) Multimethodology: Towards a Framework for Mixing Methodologies Omega 25(5)489–509.
13. Munro I, Mingers J (2002) The use of multimethodology in practice - results of a survey of practitioners Journal of the Operational Research Society 53:369–378.
14. Pidd M (2004) Modelling - Theory and Practice. Chichester.
15. Reagan-Cirincione P, Schumann S, Richardson G.P, Dorf S.A (1991) Decision Modeling: Tools for Strategic Thinking Interfaces 21(6):52–65.
16. Rosenhead J (1996) What's the Problem? An Introduction to Problem Structuring Methods Interfaces, 26(6):117–131.
17. Rosenhead J, Mingers J (2001) A New Paradigm of Analysis. Rational Analysis for a Problematic World Revisited In: Rosenhead J, Mingers J.
18. Rouwette E, Größler A, Vennix J (2004) Exploring the Influencing Factors on Rationality: a Literature Review of Dynamic Decision-Making Studies in System Dynamics Systems Research & Behavioral Science 21(4):351–370.
19. Senge P (1994) The Fifth Discipline. New York.
20. Snabe B, Größler A (2006) System Dynamics Modeling for Strategy Implementation Systems Research & Behavioral Science 23(4): July/August 2006
21. Sterman JD (2000) Business Dynamics - Systems Thinking and Modeling for a Complex World. Boston.