E. Bjørndal
M. Bjørndal
P. M. Pardalos
M. Rönnqvist
Editors

ENERGY SYSTEMS

Energy, Natural Resources and Environmental Economics

Springer

Energy Systems

Series Editor:

Panos M. Pardalos, University of Florida, USA

For further volumes:
http://www.springer.com/series/8368

Endre Bjørndal • Mette Bjørndal
Panos M. Pardalos • Mikael Rönnqvist
Editors

Energy, Natural Resources and Environmental Economics

Springer

Editors
Professor Endre Bjørndal
Department of Accounting, Auditing
and Law
Norwegian School of Economics
and Business Administration (NHH)
Helleveien 30
5045 Bergen
Norway
Endre.bjorndal@nhh.no

Professor Panos M. Pardalos
Department of Industrial & Systems
Engineering
Center for Applied
Optimization, University of Florida
Weil Hall 303
P.O. Box 116595 Gainesville
FL 32611-6595
USA
Pardalos@ufl.edu

Professor Mette Bjørndal
Department of Finance
and Management Science
Norwegian School of Economics
and Business Administration (NHH)
Helleveien 30
5045 Bergen
Norway
Mette.bjorndal@nhh.no

Professor Mikael Rönnqvist
Department of Finance
and Management Science
Norwegian School of Economics
and Business Administration (NHH)
Helleveien 30
5045 Bergen
Norway
Mikael.ronnqvist@nhh.no

ISSN 1867-8998 e-ISSN 1867-9005
ISBN 978-3-642-12066-4 e-ISBN 978-3-642-12067-1
DOI 10.1007/978-3-642-12067-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010931834

© Springer-Verlag Berlin Heidelberg 2010
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.
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.

Cover illustration: Cover art entitled "WOOD COLORS IN MOTION" is designed by Elias Tyligadas.

Cover design: SPi Publisher Services

Printed on acid-free paper

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

Preface

This book consists of a collection of articles describing the emerging and integrated area of Energy, Natural Resources and Environmental Economics. A majority of the authors are researchers doing applied work in economics, finance, and management science and are based in the Nordic countries. These countries have a long tradition of managing natural resources. Many of the applications are therefore founded on such examples.

The book contents are based on a workshop that took place during May 15–16, 2008 in Bergen, Norway. The aim of the workshop was to create a meeting place for researchers who are active in the area of Energy, Natural Resource, and Environmental Economics, and at the same time celebrate Professor Kurt Jörnsten's 60th birthday.

The book is divided into four parts. The first part considers petroleum and natural gas applications, taking up topics ranging from the management of incomes and reserves to market modeling and value chain optimization. The second and most extensive part studies applications from electricity markets, including analyses of market prices, risk management, various optimization problems, electricity market design, and regulation. The third part describes different applications in logistics and management of natural resources. Finally, the fourth part covers more general problems and methods arising within the area.

The compiled set of 29 papers attempts to provide readers with significant contributions in each of the areas. The articles are of two types, the first being general overviews of specific central subject areas, and the second being more oriented towards applied research. This hopefully makes the book interesting for researchers already active in research related to energy, natural resources, and environmental economics, as well as graduate students.

We acknowledge the valuable contributions from the Norwegian School of Economics and Business Administration (NHH) and the Institute for Research in Economics and Business Administration (SNF). We are also very grateful to all the referees and to Ph.D. student Victoria Gribkovskaia for her work on the manuscript.

Bergen/Gainesville
December 2009

Endre Bjørndal
Mette Bjørndal
Panos Pardalos
Mikael Rönnqvist

Contents

Part II Electricity Markets and Regulation

Overview of the Contributions

Part I: Petroleum and Natural Gas

Sovereign wealth funds (SWF) is the new name for assets held by governments in another country's currency. These funds are growing at an unprecedented rate and are becoming important players in global financial markets. *Døskeland* describes these funds and classifies different investment strategies.

Osmundsen discusses challenges, incentives, and ownership of petroleum reserves. The issues are discussed in relation to two cases taken from Russia and Brazil.

Mohn describes how predictions from a geophysical approach to oil exploration and production suggests that oil production will develop according to a predetermined and inflexible bell-shaped trajectory, quite independent of variables relating to technological development, economics, and policy.

Nygreen and Haugen discuss applications of mathematical programming tools and techniques in field development planning for the Norwegian continental shelf.

Midthun and Tomasgard provide an overview of the natural gas value chain, modelling aspects and special properties of pipeline networks that provide challenges when doing economic analyses.

Mathiesen describes equilibrium models to analyze the European Market for Natural Gas.

Fuglseth and Grønhaug describe how equilibrium models can enhance managerial team learning in complex and ever-changing situations.

Bengtsson and Nonås survey the planning and scheduling of refinery activities.

The focus is on identification of problems, models, and computational difficulties introduced by the models.

Part II: Electricity Markets and Regulation

Andersson and Lillestøl exploit multivariate and functional data techniques to capture important features concerning the time dynamics of hourly day-ahead electricity prices at Nordpool.

Electricity is a non-storable commodity and electricity prices follow fairly regular fluctuations in demand, stemming from time dependent variations in economic activity and weather conditions. However, it is possible to store electricity as a different energy carrier. These aspects are described by *Gjølberg*.

Bjerksund, Rasmussen, and Stensland analyze valuation and risk management in the Norwegian electricity market.

Kristoffersen and Fleten provide an overview of stochastic programming models in short-term power generation scheduling and bidding.

Chabar, Granville, Pereira, and Iliadis present a decision support system that determines the optimal dispatch strategy of thermal power plants while considering the particular specifications of fuel supply agreements.

Rebennack, Kallrath, and Pardalos discuss a portfolio optimization problem occurring in the energy market where energy distributing public services have to decide how much of the requested energy demand has to be produced in their own power plant, and which complementary amount has to be bought from the spot market and from load following contracts.

Bergendahl investigates the advantages of investing in plants for cogeneration, i.e., Combined Heat and Power (CHP), in case the heat is utilized for district heating. A focus is set on Swedish municipalities where these are an important part of energy production.

Bjørndal, Jörnsten, and Rud describe a price adjustment procedure based on capacity charges for managing transmission constraints in electricity networks.

Agrell and Bogetoft analyze electricity distribution system operators and particular challenges in the Nordic countries.

Bjørndal, Bjørndal, and Fange provide an overview of the Norwegian regulation of electricity networks after the Energy Act of 1990. Various data envelopment analysis (DEA) models are discussed.

Jennergren discusses elementary properties of allowed depreciation and return on the asset base for a regulated company under two regulatory models, the traditional rate-of-return model and the more recent long run incremental cost (LRIC) model.

Part III: Natural Resources and Logistics

Sandal and Steinshamn examine harvesting of fish in predator–prey biological models. In particular, they study whether the prey can be rescued by harvesting the predator.

Greve, Golombek, and Harris study the Norwegian pulp and paper mills and describe how they can reduce pollution and how this relates to absorptive capacity and social capital.

D'Amours and Rönnqvist describe and discuss important issues in collaborative logistics.

Edwards discusses an assignment problem where pilots are assigned to ships in the sea of Bothnia.

Andersson, Christiansen, and Fagerholt discuss transportation planning and inventory management in the LNG supply chain. They also suggest models for two typical problem formulations.

Part IV: General Problems and Methods

Hamacher and Jörnsten present a combinatorial optimization model for the relinquishment of petroleum licenses on the Norwegian continental shelf. This work has not been published earlier but forms a basis for k-cardinality tree problems.

Haugland presents an overview of models and solution methods for pooling problems.

Flåm presents a theoretical foundation including properties for cooperation under ambiguity.

Aase studies the pricing of an American put option when the underlying assets pay no dividend.

Ståhl describes applications of discrete event simulation in the area covered in the book. In particular, he discusses project management, bidding of oil resources and game with duopolies.

List of Contributors

Knut K. Aase Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Centre of Mathematics for Applications (CMA), University of Oslo, Oslo, Norway, knut.aase@nhh.no

Per J. Agrell Louvain School of Management and CORE, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium, per.agrell@uclouvain.be

Henrik Andersson Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway, henrik.andersson@iot.ntnu.no

Jonas Andersson Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, jonas.andersson@nhh.no

Jens Bengtsson Department of Accounting, Auditing and Law, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, jens.bengtsson@nhh.no

Göran Bergendahl School of Business, Economics, and Law, University of Gothenburg, SE 405 30 Gothenburg, Sweden, goran.bergendahl@handels.gu.se

Petter Bjerksund Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, petter.bjerksund@nhh.no

Endre Bjørndal Department of Accounting, Auditing and Law, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, endre.bjorndal@nhh.no

Mette Bjørndal Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Østfold University College, 1757 Halden, Norway, mette.bjorndal@nhh.no

Peter Bogetoft Department of Economics, Copenhagen Business School, 2000 Frederiksberg, Denmark, pb.eco@cbs.dk

Raphael Martins Chabar PSR, Rio de Janeiro, RJ, Brazil, chabar@psr-inc.com

Marielle Christiansen Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway, mc@iot.ntnu.no

Sophie D'Amours FORAC-CIRRELT, Université Laval, QC, Canada, sophie.DAmours@forac.ulaval.ca

Trond Døskeland Department of Accounting, Auditing and Law, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, trond.doskeland@nhh.no

Henrik Edwards Vectura Consulting AB, Box 46, 17111 Solna, Sweden, henrik.edwards@vectura.se

Kjetil Fagerholt Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway, kjetil.fagerholt@iot.ntnu.no

Kari-Anne Fange Department of Business, Languages and Social Sciences, Østfold University College, 1757 Halden, Norway, kari.a.fange@hiof.no

Sjur Didrik Flåm Department of Economics, University of Bergen, 5020 Bergen, Norway, sjur.flaam@econ.uib.no

Stein-Erik Fleten Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway, stein-erik.fleten@iot.ntnu.no

Anna Mette Fuglseth Department of Strategy and Management, Norwegian School of Economics and Business Administration (NHH), Breiviken 40, 4045 Bergen, Norway, anna.mette.fuglseth@nhh.no

Ole Gjølberg Department of Economics and Resource Management, UMB, Taarnbygningen, 1432 Aas, Norway
and
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, ole.gjolberg@umb.no

Sergio Granville PSR, Rio de Janeiro, RJ, Brazil, granville@psr-inc.com

Arent Greve Department of Strategy and Management, Norwegian School of Economics and Business Administration (NHH), Breiviksveien 40, 5045 Bergen, Norway
and
Fakultet for økonomi og samfunnskunnskap, Universitetet i Agder, Kristiansand, Norway, arent.greve@nhh.no

Kjell Grønhaug Department of Strategy and Management, Norwegian School of Economics and Business Administration (NHH), Breiviken 40, 4045 Bergen, Norway, kjell.gronhaug@nhh.no

Horst W. Hamacher Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany, hamacher@mathematik.uni-kl.de

Kjetil Haugen Molde University College, Box 2110, 6402 Molde, Norway, kjetil.haugen@himolde.no

Dag Haugland Department of Informatics, University of Bergen, 5020 Bergen, Norway, dag.haugland@ii.uib.no

Niko A. Iliadis EnerCoRD, Athens, Greece, niko.iliadis@enercord.com

L. Peter Jennergren Department of Accounting, Stockholm School of Economics, 11383 Stockholm, Sweden, peter.jennergren@hhs.se

Kurt Jörnsten Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, kurt.jornsten@nhh.no

Josef Kallrath Department of Astronomy, University of Florida, Weil Hall 303, P.O. Box 116595 Gainesville, FL 32611-6595, USA, kallrath@astro.ufl.edu

Trine Krogh Kristoffersen Risø National Laboratory for Sustainable Energy, Technical University of Denmark, 4000 Roskilde, Denmark, trkr@risoe.dk

Jostein Lillestøl Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, jostein.lillestol@nhh.no

Lars Mathiesen Department of Economics, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, lars.mathiesen@nhh.no

Kjetil T. Midthun Department of Applied Economics, SINTEF Technology and Society, 7036 Trondheim, kjetil.midthun@sintef.no

Klaus Mohn StatoilHydro (E&P Norway), 4036 Stavanger, Norway
and
Department of Economics and Business Administration, University of Stavanger, 4035 Stavanger, Norway, kmohn@statoil.com

Sigrid-Lise Nonås Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, sigrid-lise.nonas@nhh.no

Bjørn Nygreen Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway, bjorn.nygreen@iot.ntnu.no

Petter Osmundsen Department of Industrial Economics and Risk Management, University of Stavanger, 4036 Stavanger, Norway
and
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, Petter.Osmundsen@uis.no

Panos M. Pardalos Department of Industrial & Systems Engineering, Center for Applied Optimization, University of Florida, Weil Hall 303, P.O. Box 116595 Gainesville, FL 32611-6595, USA, pardalos@ufl.edu

Mario Veiga F. Pereira PSR, Rio de Janeiro, Brasil, mario@psr-inc.com

Heine Rasmussen Statkraft, 0216 Oslo, Norway, heine.rasmussen@statkraft.com

Steffen Rebennack Department of Industrial & Systems Engineering, Center for Applied Optimization, University of Florida, Weil Hall 303, P.O. Box 116595 Gainesville, FL 32611-6595, USA, steffen@ufl.edu

Mikael Rönnqvist Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, mikael.ronnqvist@nhh.no

Linda Rud Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Institute for Research in Economics and Business Administration, Breiviksveien 40, 5045 Bergen, Norway, linda.rud@nhh.no

Leif K. Sandal Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, leif.sandal@nhh.no

Ingolf Ståhl Center for Economic Statistics, Stockholm School of Economics, 11383 Stockholm, Sweden, ingolf.stahl@hhs.se

Stein I. Steinshamn Institute for Research in Economics and Business Administration, Breiviksveien 40, 5045 Bergen, Norway, stein.steinshamn@snf.no

Gunnar Stensland Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway, gunnar.stensland@nhh.no

Asgeir Tomasgard Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, 7491 Trondheim, Norway, asgeir.tomasgard@iot.ntnu.no

Part I
Petroleum and Natural Gas

Investment Strategy of Sovereign Wealth Funds

Trond Døskeland

Abstract Sovereign wealth funds (SWF) are a new name for assets held by governments in another country's currency. These funds are growing at an unprecedented rate and are becoming important players in global financial markets. In this paper, I describe how these funds are being invested and I develop a classification of investment options available for sovereign wealth funds.

1 Introduction

When a country exports more than it imports the country accumulates assets. Such a trade surplus may arise from several factors, for example, increased productivity, or new access to valuable natural resources. Over the past decade we have seen historically large financial imbalances around the globe with many oil-producing and some Asian countries running large trade surpluses on a sustained basis.[1] As accumulated reserves in these countries are well beyond the requirements for exchange-rate management, their financial leaders have started to rethink how best to manage their accumulated reserves. Many countries already have set up their own long-term investment vehicles funded by foreign-exchange assets. Other nations will surely follow this pattern. These investment vehicles have recently been named sovereign wealth funds (SWFs).

SWFs are not a new phenomenon, but in recent years, wealth accumulated in existing funds has exploded, and many new funds have been created.

[1] The balance of trade is the difference between a nation's imports and exports. The balance of trade forms part of the current account, which also includes income from the international investment position as well as international aid. If a government is going to accumulate assets in its Sovereign Wealth Fund, the government should also run a surplus on the government budget. However, there is a high correlation between trade surplus and government budget surplus.

T. Døskeland
Department of Accounting, Auditing and Law, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: trond.doskeland@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_1,
© Springer-Verlag Berlin Heidelberg 2010

The International Monetary Fund (IMF) estimated in September 2007 that sovereign wealth funds control as much as $3 trillion. This number can jump to $12 trillion by 2012.

The size and growth of SWFs raise the issue of the expanded role of governments in the ownership and management of international assets. It calls into question basic assumptions about the structure and functioning of our national economies, global investment, and the international financial system. Traditionally, in a market-based economy and financial system, the role of government is limited in the economic and financial systems. But economic and financial forces are shifting wealth toward countries with innovative conceptions of the role of government in their economic and financial systems. As governmental roles in investments change, some, for example, financial experts and government leaders, are concerned about how SWFs will be used. Will governments use SWFs simply as financial tools and eye investments from a purely financial standpoint, or will SWFs emerge as an implement of political muscle? Such a concern is expressed, for example from the United States, where foreign governments or government-controlled entities have bought large, even controlling, stakes in financial institutions. American experts wonder about the consequences of the latest bailouts of the largest US financial institutions such as Citigroup and Morgan Stanley. We might also ask what would have happened with the financial institutions during the sub-prime crisis if the SWFs had not helped. In light of the recent developments, the IMF, in close partnership with SWFs, is currently working on establishing standards for the best use of SWFs in global investment.

In the next section of the paper, I will give an overview of the development of sovereign wealth funds. In Sect. 3, I will elaborate on the investment strategy of sovereign wealth funds. Among other things, I will discuss different roles the government may have in SWFs. I will conclude with a few remarks on SWFs and the possibility of a unified theory of investment strategy.

2 The Development of Sovereign Wealth Funds

The emergence of SWFs has been a direct consequence of the rapid growth of central bank reserves. As central bank reserves in a number of countries have grown in recent years, it became apparent that they exceeded by a large margin the level of reserves necessary to ensure the precautionary objective of insulating those countries' currencies from speculative attacks.

Broadly we can divide the origin of a country's large foreign exchange reserves into two sources.

- *Commodity*. The source of the surplus is through commodity exports (either owned or taxed by the government). These are typically oil and gas, but could also be metals. 64% of SWFs have their funding source from commodities, mainly oil and gas (based on numbers from Sovereign Wealth Fund Institute).

- *Traditional trade*. Large current account surpluses have enabled non-commodity exporters (particularly in Asia) to accumulate foreign exchange reserves. 36% of assets of SWFs are funded by traditional trade.

Many of the funds are funded by a persistently large US current account deficit. In general, Asian and oil producing nations have the largest cumulative reserves, with China, Russia, and Middle Eastern countries being the fastest accumulators over the past years.

Regardless of the source of funds, all countries need some foreign exchange reserves. When a country, by running a current account surplus, accumulates more reserves than it needs for immediate purposes, it can create a sovereign fund to manage those "extra" resources. A sovereign wealth fund is often created when a country sees that it has more foreign exchange reserves than needed for risk management. Accordingly, we can divide the foreign exchange reserves into two types of reserves.

- *Risk management funds*. The funds' objective is primary stabilization. This can be the safety and liquidity of the currency or to insulate the budget and the broader economy from excess volatility, inflation and other macroeconomic threats. The funds are not set up to deliver investment returns.
- *Sovereign wealth funds*. The funds can be similar to pension or endowment funds. SWFs have a very long, often multi-decade, investment horizon and limited liquidity needs. The funds objective is long-term return and wealth maximization. We consider the investment strategy of these funds in the next section.

There is no single, universally accepted definition of a SWF. Based on the previous classification of two types of reserves, we use the term SWF to mean a government investment vehicle which is funded by foreign exchange assets, and which manages these assets with a long horizon separately from the official reserves of the monetary authorities (e.g., central bank). The assets may be managed directly by a government entity or may be subcontracted to a private entity inside or outside the country.

Estimates of foreign assets held by sovereigns include $6 trillion of international reserves (risk management funds) and about $3 trillion in types of sovereign wealth fund arrangements. It is often difficult to classify a fund as either a risk management or a SWF. Many of the funds are a combination of those two types.[2] Assets under management of mature market institutional investors are estimated to $53 trillion; about 20 times more than the size of SWFs. Hedge funds manage about $1–$1.5 trillion, modestly less than those SWFs (IMF 2007). IMF projections suggest that sovereigns (predominantly emerging markets) will continue to accumulate international assets at the rate of almost $1 trillion per year, which could bring the aggregate foreign assets under sovereign management to about $12 trillion by 2012.

[2] A SWF or risk management funds are not the only way a country can hold money. A country can also hold/own public pension funds, and state-owned enterprises. Public pension funds hold the funds that states promise their citizens. These funds have traditionally kept low exposure to foreign assets. State-owned enterprises are companies fully or partly managed by the state, each of which may have its own assets and investments.

For Asian emerging markets in particular, much will depend on how successful the countries are in controlling growth. For a SWF with asset accumulation due to oil revenues, future size is largely dependent upon the price of oil and the ability to exploit old and find new oil fields.

SWFs use a variety of disclosure and transparency standards. By this we mean that financial reporting and information about the funds vary from country to country. As a result, precise data on the current size of SWFs are hard to come by. Table 1 shows an overview of the largest funds made by SWF Institute. Some SWFs have a very long history. One of the first was the Kuwait Investment Board, a commodity SWF created in 1953 from oil revenues before Kuwait gained independence from Great Britain. As we can see in Table 1, other funds have been newly created. Twenty new SWFs have been created in the past eight years. In this period the assets under management of SWFs have grown from several hundred billions to trillions of US dollars. Currently, about 35 countries have sovereign wealth funds. Many other countries have expressed interest in establishing their own. Still, the holdings remain quite concentrated, with the top six funds accounting for 73% of total assets. The Abu Dhabi Investment Authority is the world's largest fund. The six biggest funds are sponsored by the United Arab Emirates (UAE), Norway, Singapore, Saudi Arabia, Kuwait, and China.

In Sect. 3 I outline a framework for how the SWFs should invest, and try to compare this with how they actually invest. An accurate description of current

Table 1 Sovereign wealth funds

Country	Fund name	Assets $ Billion	Inception	Source
UAE – Abu Dhabi	Abu Dhabi investment authority	875	1976	Oil
Norway	Government pension fund – global	380	1990	Oil
Singapore	Investment corporation	330	1981	Non-commodity
Saudi Arabia	Various funds	300	NA	Oil
Kuwait	Kuwait investment authority	250	1953	Oil
China	China investment company ltd.	200	2007	Non-commodity
China – Hong Kong	Monetary authority investment portfolio	163	1998	Non-commodity
Singapore	Temasek holdings	159	1974	Non-commodity
Australia	Australian future fund	61	2004	Non-commodity
Qatar	Qatar investment authority	60	2000	Oil
Libya	Reserve fund	50	NA	Oil
Algeria	Revenue regulation fund	43	2000	Oil
US – Alaska	Alaska permanent fund	39.8	1976	Oil
Russia	National welfare fund	32	2003	Oil

(continued)

Table 1 (continued)

Country	Fund name	Assets $ Billion	Inception	Source
Ireland	National pensions reserve fund	30.8	2001	Non-commodity
Brunei	Brunei investment agency	30	1983	Oil
South Korea	Korea investment corporation	30	2005	Non-commodity
Malaysia	Khazanah nasional BHD	25.7	1993	Non-commodity
Kazakhstan	Kazakhstan national fund	21.5	2000	Oil, gas, metals
Canada	Alberta's heritage fund	16.6	1976	Oil
US – New Mexico	Investment office trust funds	16	1958	Non-commodity
Chile	Economic and social stabilization fund	15.5	2007	Copper
Taiwan	National stabilisation fund	15	2000	Non-commodity
New Zealand	New Zealand superannuation fund	13.8	2003	Non-commodity
Iran	Oil stabilisation fund	12.9	1999	Oil
Nigeria	Excess crude account	11.0	2004	Oil
Botswana	Pula fund	6.9	1993	Diamonds et al.
US – Wyoming	Permanent Wyoming mineral trust fund	3.7	1974	Minerals
US – Alabama	Alabama trust fund	3.1	1986	Natural gas
Azerbaijan	State oil fund	2.5	1999	Oil
Vietnam	State capital investment corporation	2.1	2006	Non-Commodity
East Timor	Timor-Leste petroleum fund	2	2005	Oil, gas
Oman	State general reserve fund	2	1980	Oil, gas
UAE – Ras Al Khaimah	RAK investment authority	1.2	2005	Oil
Venezuela	FIEM	0.73	1998	Oil
Trinidad and Tobago	Revenue stabilisation fund	0.46	2000	Gas
Kiribati	Revenue stabilisation fund	0.4	1956	Phosphates
Uganda	Poverty action fund	0.35	1998	Foreign aid
Mauritania	National fund for hydrocarbon reserves	0.2	2006	Oil, gas
Angola	Reserve fund for oil	0.2	2007	Oil
Total		3207		

Data on assets under management reflect latest available figures as reported by each individual entity or other authoritative sources. Updated March 19 2008 by SWF Institute

investments practices for SWFs is difficult to establish. This is because of the low transparency of SWFs. Determination of their size, their investment strategies, and assessing whether SWF investments may have been shaped by political objectives, each pose special problems for the researcher.

3 Investment Strategy

SWFs are as diverse in their investment strategies as in every other characteristic. This study often uses the case of Norway as best practice, but as we will see SWFs have not yet reached a consensus on the optimal investment strategy. In this section, I will develop a framework that may help countries decide their investment strategy.

3.1 Framework for Optimal Investment Strategy

I have mentioned that sovereigns have a long investment horizon and limited liquidity needs. Often SWFs aim to meet long-term real returns objectives and can accept short-term volatility in returns for expected higher long-term returns. The funds may often gain diversification benefits from a less-constrained asset allocation. Relative to other institutional investors, SWFs have a stable funding base and no regulatory requirements or capital adequacy. One way of formalizing these properties of the investor is to formulate an optimization problem.

Primary objective is high stable long-term return and wealth maximization. The sovereign maximizes its surplus wealth, defined as W. The country uses its financial assets, defined as *FA*, to maximize wealth. As illustrated in Table 2, the relation between financial assets and wealth is restricted by a *liability profile*, defined as L. Defining the liability profile is difficult yet essential for the investment strategy. A liability is the present value of future negative cash flows. One often has to consider the broader national agenda, which could include various social, political, intergenerational and environmental liabilities. For example, environmental problems, future pensions expenditures or infrastructure, could be future liabilities for a country. Thus, the relation between financial assets, liabilities and surplus wealth is given by the following relation, $FA - L = W$. Our decision variables are related to financial assets. Our first choice is to decide which assets to invest in and then optimize the shares in the different asset classes (*asset allocation policy*). The second choice is to choose an investment style. Either the investor believes in his ability to outperform the overall market by picking stocks he believes will perform well (*strategic/active investment style*) or he may think that investing in a market index may produce potentially higher long-term returns (*passive investment style*).

Based on the written outline of an optimization problem, the rest of this section will more thoroughly examine three key factors:

- The liability profile;
- The choice of asset allocation policy; and
- The choice between passive or strategical investment style.

Table 2 The balance sheet of the country

Balance sheet of SWF	
Financial Assets (*FA*)	Liabilities (L)
	Surplus wealth (W)

For a more explicit quantitative solution of a similar optimization problem, I refer to Døskeland (2007).

3.1.1 The Liability Profile

Defining a liability profile is essential for the investment strategy. It appears, however, that investors and some sovereign wealth funds do not have defined liabilities in their strategy. Such an *asset-only* investor can be illustrated with help of Table 2. If the sovereign does not take its liability profile into account, we can assume they set $L = 0$. The surplus wealth (W) on the right-hand side will then be equal the financial assets (FA). If the fund behaves as an asset-only investor, the optimization problem of the investor is equal to a multi-period mean–variance portfolio choice problem. If we assume no time-varying investment opportunities, the asset allocation will be constant over time. The problem then collapses to a standard mean-variance problem first solved by Markowitz (1952).

For a SWF this is a too simplified framework. The fund has liabilities. There are some negative cash flows in the future the fund has to pay. Therefore, the rest of this paper will investigate the more realistic case where the investor is an *asset-liability* investor who takes its liabilities into account. The liability profile is dependent on the withdrawal rules regarding the fund's future cash flows. For traditional funds (i.e., pension plans, insurance companies, endowments, etc.) it can be known with a high degree of confidence for what purpose, when and how much money will be required. For a SWF, it is harder to identify the liability profile. Another difference is the investment horizon. While ordinary pension funds face the challenge of balancing between short-term solvency risk and long-term continuity and sustainability, the SWFs can focus on the latter. By defining the liability profile, the financial assets will be "earmarked". This has the advantage of transparency and control of the fund.

Different liability profiles characterize different types of funds and influence how they might be structured. Some SWFs combine several features in one entity. For example, Norway combines elements of stabilization, sterilization and pension reserve function. However, in principle, different liability profiles should result in different entities and structures. In the next subsection, we will see how the liability profile influences the risk-return profiles, that is, the asset allocation policy.

3.1.2 Asset Allocation Policy

The main choices in a sovereign's asset allocation policy are the selection of asset classes and their weights. Based on this process the fund will define a benchmark portfolio. This is a portfolio against which the performance of the fund can be measured. It is not easy to find information about the asset allocation policy for different SWFs. In Table 3, I have listed the available information for 10 of the largest SWFs. Only for Norway and Australia it is possible to identify the asset allocation. It seems

Table 3 Investment strategy and asset allocation of the 10 largest SWFs

Country and fund Name	Assets $ Billion	Investment strategy and strategic asset allocation
UAE – Abu Dhabi Abu Dhabi investment authority	875	Investment strategy and asset allocation is unknown
Norway government pension fund-global	380	Defined strategy. Global asset allocation with 60% in equities and 40% in global fixed income
Singapore investment corporation	330	Global asset allocation (not made public). Invests in all major asset classes
Saudi Arabia various funds	300	Major global investor. The investment strategy and asset allocation is not known beyond broad indications.
Kuwait – Kuwait investment authority	250	Two funds, both with known strategy. One of them holds 60% in MSCI-stocks, private equity and real estate.
China – China investment company ltd.	200	Not available
China – Hong Kong monetary authority investment	163	Not available
Singapore temasek holdings	159	Asset allocation unknown, but geographical distribution disclosed (38% in Singapore, 40% in rest of Asia)
Australia Australian future fund	61	Will implant international best practice for institutional investment
Qatar – Qatar investment authority	60	Asset allocation undisclosed. Holds significant stakes in foreign companies, participates in buyouts.

there are still many funds that do not have a defined asset allocation policy and a corresponding benchmark. Despite the lack of benchmarks, we see two trends in the asset allocation of the SWFs. First, an increasing number of SWFs switch from neutral assets like US Treasury bonds to a more diversified investment portfolio, with a higher level of risk accepted in search of higher returns. The second trend is a move from traditional assets like stocks, governments bonds and T-bills to other asset classes such as commodities, hedge funds, private equity, infrastructure, and real estate. For example, Norway recently announced that it will invest 5% of its fund in real estate.

As mentioned, many sovereigns do not define an asset allocation policy, and they probably do not link the asset allocation policy to the liability profile. I will now propose a framework for how sovereigns might develop an asset allocation policy and a corresponding benchmark. We find the benchmark by identifying the connection between the liabilities and the asset classes. Our objective is to have as large as possible wealth (W from Table 2) with lowest variance. We use the financial asset to neutralize (hedge) the liabilities. Therefore, it is essential to understand the long-term relation between the liabilities and the financial assets. Financial assets

and liabilities have to be characterized by the same parameters (expected return, variance and covariance) to fit in the risk return space. This is done by replication of investable assets that are as closely correlated as possible to the liability. The risk of any asset class with return R is measured with its covariance with the wealth:

$$\mathrm{Cov}(R, W) = \mathrm{Cov}(R, FA) - \mathrm{Cov}(R, L).$$

Based on the relations from the balance sheet, we can split the covariance into two parts. The first term is the covariance between the asset class with return R and the rest of the financial assets. The second term is the covariance between the same asset class and the liabilities. If the asset class covaries negatively with the value of the financial asset and/or positively with the value of the liability, the asset class hedges the wealth. The asset class should then be rewarded with a large share of the financial assets.

Traditional measures of the covariance are related to short-term variability in market value (e.g., contemporaneous correlation). For a long-term investor with long-term liabilities, this is not appropriate because the long duration dimension is not captured. For example, by a contemporaneous correlation it may look like the correlation between stocks and wages is low, if so, stocks are not a good hedge for pensions (related to wages). But using other methods to capture the long-term relation (e.g., cointegration and duration matching), I find that there actually is a positive relation. This implies that pension funds (and those SWFs with pensions in their liability profile) should invest more heavily in stocks. Døskeland (2007) looks at investment strategy for sovereign wealth funds with Norway as a case. He finds that stocks have a more positive relation than long-term bonds to pensions, implying a high stock share in the benchmark of the Norwegian Pension Fund. Thus, if Norway was to take into account the long-term relation between financial assets and pension liabilities, Norway should have a high fraction of stocks in their fund. Recently, Norway has increased its share in equities from 40% to 60%.

The asset allocation policy should reflect the long-term relationships between the financial assets and the liabilities. The fund should invest in a broadly diversified portfolio, investing a major portion in what would be traditionally viewed as "risky" assets, primary stocks. In our case, investing primarily in risk-free or "safe" assets could be worse than investing in the so-called "risky" assets – it could effectively guarantee that the fund would not be able to meet its sizable liability. But the funds should also consider other classes such as private equity, corporate bonds, emerging markets, real estate, venture capital, infrastructure funds, and hedge funds. The funds should not invest in the same commodity or commodity companies as their largest revenue of the home country. It is strange that Norway, whose main revenue comes from oil, invests large fractions of their portfolio in oil companies.

Whether the liabilities are real or nominal influences the asset allocation policy. For the risk management fund the financial assets often are invested nominally because when a sovereign needs the money they are needed nominally. For the more long-term SWF the situation is different. An endowment fund typically needs to maintain a certain amount of annual spending, while at the same time preserving

the real value of the principal to continue paying annual amounts in the future. A pension fund's liabilities are also real, in that it will be making payments to retirees based on some formula which includes real incomes.

The perfect match for real liabilities is inflation protected indexed linked bonds. But these instruments yield a low return. If we compare nominal bonds and stocks, we find that stocks are often a better hedge against inflation than nominal bonds. With increasing inflation, the effect tends to be passed over to the customers in the prices of goods and services. Thus, the real value of the company is quite stable. This effect was first pointed out by Bodie (1976). Thus, stocks are a good match for real liabilities.

The question whether to invest in the home country or internationally is also influenced by the liability profile. For a fund with a strong focus on pension provision, its liabilities by definition must be denominated in the currency of the owner. Yet, all of Norway's assets are invested internationally and denominated in foreign currencies. Norway is a developed country with a high per-capita GDP and a well-developed public safety net. Domestic investments could have driven down the domestic return below the international return. Domestic investments could also become subject to political processes which may reduce financial returns and transparency. Therefore, Norway has limited all of the pension fund's investments to overseas markets. The strategy also fits in with the fund's endowment approach; it is simply transforming its concentrated exposure to volatile oil into a much more balanced and diversified exposure within a broader global economy.

But Norway is a rich developed country. For an emerging market economy an implicit currency bet (liabilities denominated in the home country and investments internationally) would be likely to lose money in the long term. As emerging markets economies "catch up" in their levels of productivity and economic development, their currencies, all other things begin equal, will almost certainly experience real appreciation. It is probably not in the interest of an underdeveloped economy to suppress current consumption and capital formation by the present generation – all for the sake of maximizing financial savings of future generations. It is better for the future generations to inherit a diversified highly advanced local economy than a global financial portfolio. Thus, even if the framework for developing an asset allocation policy is quite similar, different countries have different characteristic that make the asset allocation policy heterogeneous.

3.1.3 Passive or Strategic/Active Investment Style

The final choice is to choose an investment style, either passive or strategic/active investment style. As we see in Fig. 1 along the vertical axis, the different funds have different strategies related to strategic/active and passive investments. The simplest and most efficient way of investing (if one believes in efficient market) is to mimic the benchmark defined in the asset allocation policy. Norway's SWF has almost a passive investment style. The fund follows the defined benchmark closely, but

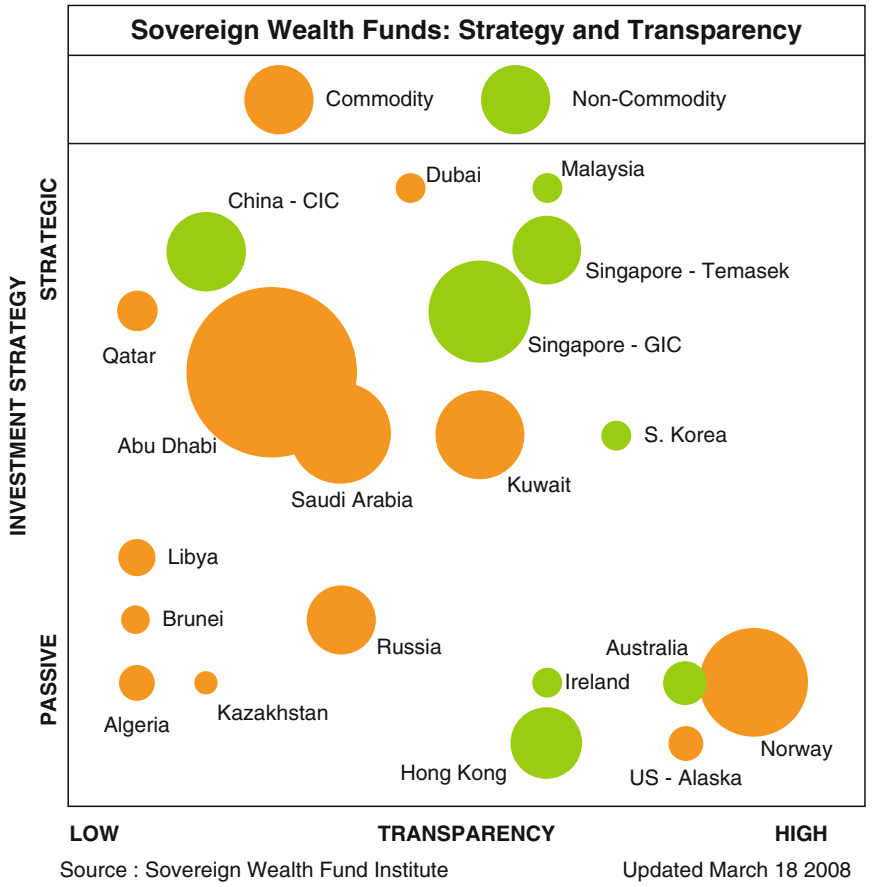


Fig. 1 Investment style and transparency of SWFs

deviates with small bets from the benchmark. The alternative investment style is to invest strategically or actively. For this alternative, the fund often does not have a defined benchmark. The overall goal is to maximize returns in absolute terms. In Fig. 1 we see that the strategic/active funds also often have lowest transparency.

For a fund with strategic/active investment style it is hard to separate whether the investments are done to directly maximize returns or are more so-called *political motivated* investments. As SWFs are foreign state-owned investment vehicles, their investments may raise concerns for the recipient country. A politically motivated investment often improves the long-term, overall economic situation for a sovereign by improving access to markets, technological advances, etc. In other words, politically motivated investments pursue dominantly indirect economic return and/or political benefits in order to increase the social welfare and/or to enlarge the nation's leeway in the international economical and political arena. Investments in so-called

key industries (defense, infrastructure, communications, financial institutions, etc.) are politically most feared. The low level of transparency in the majority of SWFs heightens the suspicions of the recipient country.

Several SWFs have gained public interest for specific investments. SWFs acquiring assets in the Unites States and elsewhere are creating concern among policymakers. Sovereign funds such as Abu Dhabi's Mubadala have invested in private equity firms, including Carlyle, the American buyout giant, and Och-Ziff, a hedge fund. Dubai's SWF has bought up shares of several Asian companies, including Sony. In Germany they are concerned that Russia is buying up pipelines and energy infrastructure and squeezing Europe for political gain.

Money from SWF funds have participated strongly in the rescue of stricken Wall Street banks after the sub-prime crises. Oil-rich nations as well as funds and banks in Asia, have injected $41 billion of the $105 billion of capital injected into major financial institutions since November 2007 (IMF 2008). Examples of strategic investments done by SWFs in the US financial sector are listed in Table 4.

4 Conclusion

The growth of SWFs has raised several issues. First, the likely growth of SWFs and the substantially higher risk tolerance of those who will be taking their asset allocation decisions may be large enough to have an impact on the average degree of risk tolerance of investors across the world. This reduces somewhat the attractiveness of relatively safe assets and increases the attractiveness of riskier. As a result of this shifting pattern of demand, we can expect that bond yields gradually rise and reduce the equity risk premium.

Second, if one believes in efficient markets, funds should not have an active investment style. However, both among academics and professionals there is no consensus about the best investment style. For example, before the financial crisis the Norwegian Government Pension Fund had a quite large extra return from active investing, but during the crisis almost all extra return is lost.

Third, many of the funds are operating as independent investment vehicles. The funds should rather be used as an integrated diversification tool. For example, the Norwegian Government Pension Fund should not invest in oil companies. One argument in favor of not imposing restriction on the investment universe is that it makes it more difficult to evaluate the fund's performance. But at the same time the Norwegian Government Pension Fund has restriction when it comes to ethical and environmental investments, thus restricting the fund from investing in oil companies should not be a problem.

There have been expressed concerns about SWFs transparency, including their size, and their investment strategies, and whether SWF investments may be affected by political objectives. Despite innumerable publications, statements and research articles there are no common agreement on the investment strategies of the SWFs. In this paper I have tried to describe some instructions for how to develop

Table 4 SWF capital injection into financial institutions

Date of Announcement	Financial Institutions	Writedown (of financial institution)	SWFs	Other investor(s)	Amount (percent of total stakes) from SWFs	Other investor(s)
November 26, 2007	Citigroup	$6 billion in 2007:Q3	Abu Dhabi investment authority	–	$7.5 billion (4.9%)	–
December 10, 2007	UBS	$18 billion in 2007	Government of Singapore investment corporation	Unknown Middle Eastern investor	$9.7 billion (10%)	$1.8 billion (2%)
December 19, 2007	Morgan Stanley	$9.4 billion in 2007:Q4	China investment corporation	–	$5 billion (9.9%)	–
December 21, 2007	Merrill Lynch	$8.4 billion in 2007:Q3	Temasek holdings	Davis Selected Advisors, L.P.	$4.4 billion (9.4%)	$1.2 billion (2.6%)
January 15, 2008	Citigroup	$18.1 billion in 2007:Q4	Government of Singapore investment corporation, Kuwait investment authority	Sanford Weill, Saudi Prince Alwaleed bin Talal, Capital Research Global Investors, Capital World Investors, New Jersey Investment Division	$6.8 billion from Government of Singapore Investment Corporation (3.7%) and $3 billion from Kuwait Investment Authority (1.6%)	$2.7 billion (1.5%)

(continued)

Table 4 (continued)

Date of Announcement	Financial Institutions	Writedown (of financial institution)	SWFs	Other investor(s)	Amount (percent of total stakes) from SWFs	Other investor(s)
January 15, 2008	Merrill Lynch	$14.1 billion in 2007:Q4	Korea Investment Corporation, Kuwait Investment Authority	Mizuho Financial Group Inc.	$2 billion (3.2%) from Korea Investment Corporation and Kuwait Investment Authority, respectively	$2.6 billion (4.1%)
February 18, 2008	Credit Suisse	$2.85 billion	Qatar investment authority	–	Approximately $500 million (1–2%); the purchase was on the open market	–

Source: IMF (2008)

a framework for the investment strategy. Only the future will show whether there will be developed a unified theory for the investment strategy of SWFs. And will the politicians keep their fingers of the table and not intervene in the decisions made by the funds?

References

Bodie, Z. (1976). Common stocks as a hedge against inflation. *Journal of Finance, 31*, 459–470.

Døskeland, T. M. (2007). Strategic asset allocation for a country: the Norwegian case. *Financial Markets and Portfolio Management, 21*(2), 167–201.

IMF (2007). *Global financial stability report. World economic and financial surveys, October*. Washington, DC: International Monetary Fund.

IMF (2008). *Global Financial Stability Report. World Economic and Financial Surveys, April*. Washington, DC: International Monetary Fund.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance, 7*(1), 77–91.

Chasing Reserves: Incentives and Ownership

Petter Osmundsen

Abstract Oil companies are concerned to replace the petroleum reserves they produce in order to maintain their future level of activity. Reserves also represent an important input when analysts value these companies. However, they have to rely on booked reserves, which are subject to measurement errors. Many producer countries want to control their own resources, a goal which can come into conflict with the desire of the international companies for booked reserves. Where oil companies do not own the reserves, they may have insufficient incentives to maximise value – harmonising goals between a resource country and an oil company can be difficult. This chapter discusses the relationship between reserves and financial incentives, and between reserves and valuation. The issues are illustrated throughout with reference to two cases: StatoilHydro's projects on Shtokman in Russia and Peregrino in Brazil.

1 Introduction

In the Norwegian petroleum tax system, ownership of the resources rests with the participants in a licence subject to conditions specified in the licence, the licence agreement and more general regulations. The state often has its own share of the licence through the State's Direct Financial Interest (SDFI), which is managed by Petoro.[1] Many would say that ownership of the resources is very important, not only for the companies but also for the resource state. This is fairly obvious in the case of the companies. Ownership makes it possible to carry the reserves on the balance sheet, which financial markets want to see. Currently, great attention is being paid

[1] www.petoro.no

P. Osmundsen
Department of Industrial Economics and Risk Management, University of Stavanger, 4036 Stavanger, Norway
and
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: Petter.Osmundsen@uis.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_2,
© Springer-Verlag Berlin Heidelberg 2010

to the reserve replacement rate (RRR) of oil companies. We see constant references, for instance, to basic valuation methods in which the value of an oil company is equated with reserves in different countries multiplied by an estimated value per barrel of oil equivalent (boe) in each country of production. For these figures to be meaningful, however, one must operate with expected rather than booked reserves.

Private ownership in the licences is important for the state because it establishes incentives to maximise value creation. Replacing ownership with other types of incentives is difficult, which presents a major challenge to producer countries where the government will not allow foreign companies to own petroleum resources for one reason or another. Ownership provides long-term incentives, where the companies wish to maximise value throughout the life cycle of the field. At the same time, achieving homogenous ownership composition in licences across field areas with reservoir contact (unitisation) is important in order to avoid sub-optimisation.

In everyday parlance, people often say that oil companies own their share of the resources in a field on the Norwegian Continental Shelf (NCS). However, this is not strictly correct. Oil companies are only *licensees*, who produce the petroleum resources on behalf of the state. Ownership of underground resources is vested in the state, which gives the government the legal authority to regulate various aspects of reservoir management. On the other hand, the licensees own and control oil and gas once they come to the surface. That ensures financial incentives to maximise the value of the resources. When regulating the oil companies, more often, the government is subject to the Act on Public Administration and the standards this sets for objectivity and orderly procedures. That is relevant at present in connection with the development of the Goliat field in the Barents Sea. When the licensees have received a production right, the government cannot refuse to allow the licence partners to develop the field (as some seem to believe). However, it can impose objectively justified and non-discriminatory requirements related to the development.

2 Booked Reserves

Since estimating actually expected cash flow for oil companies is difficult and time-consuming (due to asymmetric information between analysts and company), analysts use various indicators to make rough estimates of value. A key indicator today is the RRR. This expresses how large a proportion of production in the present year has been replaced by new booked reserves, i.e. a RRR of 1 means that the booked reserves are unchanged. The ability of the companies to maintain reserves ready for production in relation to on-going recovery says something about sustainability and growth opportunities for the company, which is clearly highly relevant for valuation. That depends, of course, on the indicator being free of measurement errors. Results from analyses we have undertaken in the Department of Industrial Economics and Risk Management at the University of Stavanger indicate that no clear relationship exists between the RRR shown in the accounts and valuation; see e.g. Misund et al. (2008). This is due to the measurement errors associated with booked reserves, which are explained below.

Still, analysts are very concerned with booked reserves. This is because booked reserves are published in the quarterly accounts, whereas expected economic reserves are not readily available to analysts. When StatoilHydro recently presented its accounts, much of the focus of the business press was on reserve replacement:

> "Another worry was Statoil's reserve replacement rate, which tracks the rate at which new discoveries offset production. The rate tumbled from 86.0% in 2007 to 34.0% in 2008. Although the three-year average rate came in at a healthier 60.0%, analysts voiced their concern over the steep decline."[2]

Analysts' focus, however, does not necessarily correspond with that of the investors, and the stock price is determined by the latter. The findings of Misund et al. (2008) are that analysts and investors do not agree when it comes to the value relevance of booked reserves. Accordingly, company managers should keep their calm and not aggressively acquire new reserves in a sellers' market, even if the RRR is taking a temporary drop. Later in this chapter I analyse two investment cases for StatoilHydro, and inquire whether they are selected on the basis of sound economic valuation or whether they are driven by an urge to satisfy analysts' demand for short-term accounting reserves.

Several factors explain the lack of correspondence between booked reserves and valuation. First, the figures on reserves comply with the conservative accounting principles of the US Securities and Exchange Commission (SEC). These involve such substantial measurement errors that they fail to provide a good expression of the actual position for reserves. Second, investors will make their own reserve estimates in any event. They are clearly not going to overlook the fact that StatoilHydro has a substantial share in the Shtokman development, for instance. Focusing on single indicators underestimates investors. They are concerned with cash flow and cannot be deceived by high figures for booked reserves.

The information value of booked reserves suffers from a number of weaknesses. Reserves are recognised on the basis of the spot oil price at the balance sheet date, which does not necessarily represent a best estimate for future oil price developments. Booked reserves do not provide a consistent picture of reserves under different contracts (an income tax system, for instance, will yield higher reserves than one based on production sharing for identical cash flows). Perhaps, the most important objection to the conservative rules, however, is that the reserve figures do not provide complete information on the subsequent growth of the company and thereby on the sustainability of its operations. This is because they do not include less mature reserves, which can vary a great deal from one company to another. In any event, the attention given to booked reserves helps make the NCS more attractive. The Norwegian licence model gives companies greater opportunities to carry reserves than is the case in nations which operate with production sharing agreements, contractor contracts and the like.

[2] http://www.forbes.com/2009/02/17/statoilhydro-bp-shell-markets-equity-0217_oil_09.html

2.1 Differences Between PSC and Concession Reserves

Traditional oilfield concession ownership is found in the OECD-area. Under this system, if producers generate a profit from on-going extraction, they pay corporation tax, sometimes supplemented with royalty or other taxes. In this instance, producers own the underlying reserves, with reported reserves being the recoverable reserves from the reservoir in total, and future physical reserve entitlement is unaffected by price volatility.

Production sharing contracts exist in many of the world's newer oil producing and non-OECD regions including West Africa, Kazakhstan, Indonesia and Egypt. The proliferation of these agreements in the 1990s has been a direct result of the government's desire to reclaim control of natural resources once a fair return has been earned by the corporate producers.

PSC agreements vary widely but typically provide oil companies with a guarantee to cover a return on their capital costs and, in exchange, impose a reserve entitlement structure. The contract generally escalates participation sharing by the local government based on the price of oil and in some cases the volume of oil pumped. As explained by Kretzschmar et al. (2007), the PSC allows contractual contingent claims (often in forms of taxation or production sharing) to be made against producer reserves when an agreed threshold of return is met and costs have been covered. This interpretation recognises the contractual nature of possible fiscal claims against oilfields (Lund 1992).

The most marked difference between concession ownership and production sharing disclosures is that reserves and production do not vary in response to oil price movements for concession fields, while both production and reserves vary under PSC regimes. For concession fields, the oil company has an equity share in the field, which does not vary with the oil price. The oil company is simply entitled to its equity share of production and reserves. Under a PSC contract, on the other hand, the oil company is to be paid a certain amount of oil to cover costs (cost oil) and profits (profit oil). When oil prices rise, the number of barrels of oil needed to pay for costs and profits are reduced. Kretzschmar et al. (2007) illustrate this with field data from the Gulf of Mexico, where reserve and production entitlement remain unchanged across the full price range USD 22.5–90. Angolan PSC reserves, by comparison, actually decrease by 0.451% per 1% oil price change in the range USD 22.5–33.75 and decrease by 0.388% in the range USD 67.5–90. Production entitlement, by comparison, also reduces in Angola, but by 0.291 and 0.181% respectively over the same price intervals. In line with Rajgopal (1999), Kretzschmar et al. (2007) recommend that supplementary information should disclose the effects of oil and gas price changes on underlying reserve disclosures.

2.2 Petroleum Reserves: Definitions

Figure 1 is a graphical representation of various definitions of petroleum reserves. The horizontal axis represents the range of uncertainty in the estimated potentially

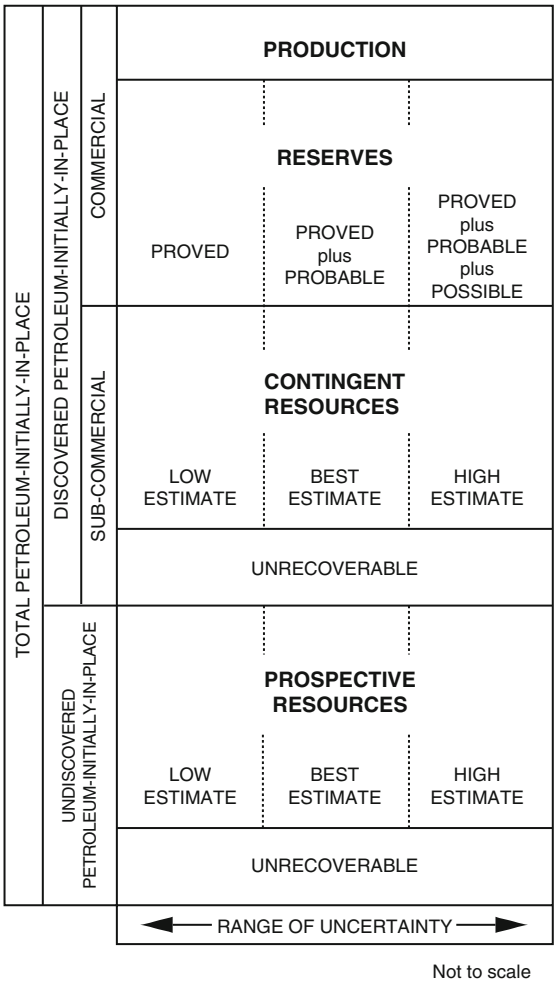


Fig. 1 Petroleum resource classification. Society of Petroleum Engineers[3]

recoverable volume for an accumulation, whereas the vertical axis represents the level of status/maturity of the accumulation.

Resources definitions vary. Some define it as including all quantities of petroleum which are estimated to be initially-in-place; however, some users consider only the estimated recoverable portion to constitute a resource. In any event, it should be understood that reserves in an *accounting* sense constitute a subset of resources, being those quantities that are discovered (i.e. in known accumulations), recoverable, commercial and remaining. This is a very conservative estimate, as it does

[3] http://www.spe.org/spe-app/spe/industry/reserves/mapping.htm

not account for contingent and prospective resources. Thus, the expected level of petroleum reserves in an *economic* sense is larger. Another distinction is the treatment of oil and gas prices. The reserve concept is always a physical volume, but the extent of reserves is contingent on the level of the oil and gas prices. At high prices, more resources are economic to extract, and reserves are accordingly higher. Whereas economic reserves are based on expected energy prices, booked reserves are based on the listed spot price at December 31.

The most widely used reserve disclosure is the one required by the SEC, owing to the importance of US capital markets and the fact that most major private oil companies have a US listing. Here companies are required to report their 'proven' reserves in a deterministic way, quite different from the probabilistic ways allowed on other exchanges.

Arnott (2004) points out that it should always be remembered that the SEC rules were introduced with the sole purpose of protecting shareholders. They were brought in at a time when most of the US oil industry was still onshore, where regular grid well-spacing was common and therefore it was fairly easy, using deterministic methods, to calculate not just the volume of remaining oil in place but also its value. However, the oil and gas industry has subsequently witnessed a major technological revolution. It is therefore ironic, according to Arnott, that at the very time that the oil and gas industry is basing more and more of its investment decisions on the results of measurements from new technologies, the SEC has tightened up its definition of what can or cannot be reported and by inference has ruled out measurements from these technologies.

The complaint about booked reserves is that this does not reflect economic reality or the reserves that the company is using when formulating its internal plans and projects. By only counting proven reserves, the SEC rules systematically understate the true extent of the resource base. Another obvious example of measurement bias is the oil produced in Canada from mining operations in tar sands. The SEC does not allow such oil to be booked as petroleum reserves on the grounds that it is a mining product – although it is at least as predictable as the oil from underground reservoirs.

Analysts are facing two problems: (1) the definition of reserves is not adequate, and (2) given the emphasis that the market is perceived to put on booked reserves, some oil companies have been tempted to manipulate their accounts – overstated the booked reserves according to the prevailing reserve definitions. Some instances of overbooking have raised uncertainty regarding booked reserves. According to Arnott (2004) the practice of 'smoothing' reserves' bookings in order to show steady reserves growth can be just as misleading to investors as overbooking.

2.3 The Role of the Reserves Report

Oil company reserves' disclosures are according to Arnott (2004) one of the most important pieces of information that the financial sector requires in order to analyse, compare and contrast the past and prospective operational performance of oil and gas exploration and production firms. Recent reserves re-categorisations by several

companies, after inquiries and questions of over-reporting from SEC, have only served to highlight the inadequacy of the published information.

Arnott states that a company's internal information structure of future production estimates is not suitable for communication outside the company for many reasons.

- It would be dynamic, complex and difficult to interpret without full knowledge of all the company's practices and parameters – in other words without being inside the company.
- It would prejudice the company in competitive bids and negotiations if this information were available to its competitors and counter-parties in negotiation.
- It is often subject to confidentiality agreements.

Obviously some communication with respect to reserves is necessary for private companies with equity or bonds held on public stock and bond markets, since:

- The expectations of future production are an important predictor of a company's future capacity to reward shareholders and repay debt-holders. Thus, accurate information is important for the companies to raise capital to new projects.
- The reported current profits depend on the allocation of exploration and developmental costs between depreciation (charged over the lifetime of production) and current expense (charged to current profits). Reported accounts therefore require a definition of expected future production – typically described as '*proven*' reserves (see Fig. 1) on a base which can be understood by investors and creditors of the company.

3 Shtokman

Ownership has been much discussed in connection with Russia's big Shtokman field in the Barents Sea. This discovery is estimated to contain a total of 3,700 billion cubic metres of gas, making it 10 times larger than the Ormen Lange field in the Norwegian Sea. StatoilHydro signed an agreement with Gazprom on 27 October 2007 concerning participation in the first phase of a Shtokman project. Gazprom, Total and StatoilHydro have concluded a shareholder agreement over the Shtokman Development AG company, which will be responsible for designing, developing, constructing, financing and utilising the facilities in a first Shtockman development phase.[4] Gazprom has 51%, Total 25% and StatoilHydro 24% of this company, which is registered in Switzerland. Total and StatoilHydro will own the phase one infrastructure for 25 years from the start of production on the field. StatoilHydro has indicated that the company's share of the gas resources corresponds to roughly 800 million barrels of oil.[5] Investment in phase one alone is likely to exceed NOK 100 billion.

[4] http://www.statoilhydro.com/no/NewsAndMedia/News/2008/Pages/ShtokmanDevelopmentAG.aspx

[5] *Dagens Næringsliv,* 22 February 2008.

Basically, the incentives in this case do not look correctly configured. The development company appears to own the infrastructure rather than the actual field. Sevmorneftegaz, a wholly owned subsidiary of Gazprom, reportedly holds the exploration and production licence for gas and condensate. The relationship between Shtokman Development and Sevmorneftegaz will build on a contract which specifies that the latter bears all financial, geological and technical risk related to production of gas and condensate and to gas liquefaction. It would thereby seem that the Russians will retain the aspects which normally fall to an oil company. OAO Gazprom owns all the shares in Sevmorneftegaz, and all the rights to market the output.

This is a contract which appears to lie closer to the type of agreement concluded by a contractor, rather than to those to which an oil company normally becomes party. Furthermore, Total and StatoilHydro only own the infrastructure for the first development stage. It is doubtful whether this provides sufficient incentive to maximise total value creation over a period of time for the whole field. This breaches elementary principles for designing incentives – a supplier should have responsibility for the areas it can affect. Knowledge of reservoir conditions represents specialist oil company expertise. Even without ownership of the actual reserves, it would have been possible to create incentives by allowing rewards to be conditional on the production portfolio.

This contract recalls contractor agreements on the NCS, where the contractor bears responsibility for delays and cost overruns but does not participate in the upside or downside related to production and gas price trends. The limited upside – which is a certain return on capital invested or a fixed sum – will often be balanced in such cases by a limited downside (both in the formulation of the contract and in its application), so that the limited opportunities for a return are proportionate to a limited risk. StatoilHydro has also concluded contractor-like contracts in Iran. These service fee deals specify what the oil company will receive, with the government taking the rest. This is the opposite of the practice in most other producer countries, where the government's share is specified and the oil company receives the residual income. Payment takes the form of oil. Cash reimbursement of costs, known as buy-back, is converted to oil at an agreed price. That makes it possible for StatoilHydro to book the reserves. The problem in this case is that the limited upside is not balanced by any downside limits. A substantial challenge has also been that the regulatory authorities, the state oil company, and supplier companies are represented by the same people and ownership. That clearly puts the foreign oil company in a weak negotiating position. Conditions in Russia are related.

Experience specifically from Iran makes it unlikely that StatoilHydro will be willing to accept a traditional contractor agreement. In this context, it is worth noting a comment from the head of the company's Moscow office, Bengt Lie Hansen: "Our exposure will be normal for an oil company – in other words, to both revenue and costs from operation of the field."[6] This must mean that Sevmorneftegaz, which has

[6] http://web3.aftenbladet.no/innenriks/okonomi/article536237.ece

been allocated all upside in the field under the terms of the shareholder agreement, will pass some of it on to the other participants. The upshot is that this will actually become something which resembles an income tax system. How far and in what way upside will be transferred to the foreign companies is unlikely to be determined until 2009. Instead of relating the incentives directly to ownership in the licence, in other words, ownership is being established in the infrastructure and efforts are being made to create synthetic incentives which will imitate the terms ordinarily enjoyed by international oil companies.

Obvious challenges here will be the credibility of the terms and the threat of renegotiation. However, it could be objected that these challenges are also present in other producer countries. Given their desire for greater predictability, the oil companies have often sought production sharing agreements because these – unlike income taxes – represent legal contracts which are more binding on the resource country. However, developments in recent years – not least in Russia – have demonstrated that production sharing agreements are incomplete contracts which give the international companies no protection worth mentioning. According to industry sources, the Russians do not want a production sharing agreement for Shtokman. Instead, they want the field to be taxed in accordance with the Russian tax regime for the petroleum sector. The exact terms will nevertheless be subject to negotiation. The Russians are likely to insist that the international participants carry the bulk of the financial risk. A normal method of doing this would be to let StatoilHydro and Total carry (pay in advance) Gazprom's development costs and pay substantial royalties charged on top of ordinary income tax regardless of the financial position of the project. With such terms, StatoilHydro and Total are guaranteed the downside in the project. The question is whether that is balanced by a corresponding and credible upside.

The decision to give Gazprom-owned Sevmorneftegaz full control of the gas resources is usually referred to as an example of the resource nationalism which is widespread in producer countries outside the OPEC area. In Russia, the starting point was a few oligarchs who had become billionaires in a very short time through unreasonably favourable asset deals with the government. A key element in Putin's agenda, which Norwegians not least must respect, was precisely that the petroleum resources should benefit the Russian people. However, the problem in Russia and many other producer countries is that a nationalistic superstructure can hinder the foreign participation needed to maximise the value of the resources for the population at large. Ownership of and control over resources are the very core of resource nationalism. Politicians in Russia could not say to their people that part of the ownership or control had been transferred to foreign companies, even though this might be just what is required by pragmatic prosperity considerations.

An article in Norwegian technical weekly *Teknisk Ukeblad* of 21 November 2007 notes that Russian legislation hinders reserves being booked on the balance sheet, and that the Russians are unlikely to amend the law simply to please the shareholders of StatoilHydro or Total. Pursuant to Russian law, Gazprom has the sole right to sell gas from Russia. This provision must be changed if StatoilHydro is to be able to carry reserves from Shtokman on its books. Such an amendment must be submitted

to the Duma (parliament), says third secretary Alexey Rybkin at the Russian embassy in Norway. The accounting rules are probably being interpreted too narrowly in this case. If StatoilHydro and Total through their participation in Shtokman secure rights to some of the production (because cost reimbursement and profits are paid in the form of gas), they can recognise the reserves even without direct ownership. This is the approach taken by StatoilHydro in Iran. What may be a bigger challenge is that resource nationalism has proved to encourage a number of populist decisions – typically a failure to respect signed agreements – which benefit neither the oil companies nor the population of the host country in the long run. In Russia, for instance, this could take the form of renegotiating terms if the project goes well and StatoilHydro and Total make money. The same willingness to renegotiate cannot be expected if project progress is poor and the companies suffer losses. An asymmetry of this kind in frame conditions clearly represents poor business economics.

> "When presenting the interim figures, [StatoilHydro CEO] Lund said that the Shtokman partnership had to be viewed in a strategic light, both because Russia is an interesting country for StatoilHydro and because the company will get the opportunity to continue the development of technology for Arctic regions."[7]

"We hope that more opportunities will open for us through this innovative contract and the special connections we have with Gazprom,"[8] said Arnaud Breuillac, the man responsible for Total's projects in central Europe and the Asian mainland.

The word 'strategic' is often used by chief executives in connection with projects which do not satisfy their company's general internal rate of return requirements. In such cases, the investment decision is based on an assessment of supplementary value, which is often relatively subjective. An example is that moving into a new area can generate additional opportunities (bridgehead investment – growth options).

Since the merger, StatoilHydro has inherited the reserve replacement challenges which faced Hydro as a separate enterprise. Like virtually all the international oil companies, it is accordingly under pressure to secure new resources. With record oil prices in 2008, a danger existed that future production was purchased at an excessive price. StatoilHydro has a balanced portfolio, where activities are spread over a number of fields in many producer countries. It has a high weighting of projects with low country risk – typically in the OECD area. This also goes for new projects. Nevertheless, it is uncertain whether increasing exposure to Shtokman makes sense in portfolio terms (risk spreading). Excessive exposure to a single project will normally be undesirable, and Russia poses a substantial country risk. Other oil companies have had their assets in Russia confiscated with little compensation, and it is difficult to find examples of oil companies who have actually made money there. The tax system is unpredictable, including uncoordinated taxation at several levels, and demands can be made to sell part of the production locally at below international

[7] DN.no, 27 February.

[8] DN.no, 19 March.

market price. In addition, main partner Gazprom – with the Russian state as its principal shareholder – is used as a political instrument. That said, the risk must be measured against the alternatives in other producer countries, which are not necessarily better. Account must also be taken of the fact that the renegotiated tax agreements in Russia were not initially framed in an optimum manner from the perspective of the Russian government. Among other facts, they were drawn up at a time when the Russian state had been weakened. The oil companies should have expected a renegotiation. Putin has also done a good deal to improve predictability in Russia, partly through greater centralisation of resource taxation.

According to press reports, Total will pay USD 800 million simply for the right to *book* reserves for Shtokman.[9] If this is correct, the Russians have understood that the oil companies' need to carry reserves on their books and that they have charged separately for this. StatoilHydro, on the other hand, is not paying anything at present. Assuming that the company has had a genuine choice in this respect, the decision to pass on recognising reserves appears basically sensible.[10] The different strategies pursued by the two companies relate to their need to make themselves attractive to investors. All companies want to present accounts which ensure the highest possible market valuation. When Total pays USD 800 million for its 25% of the Shtokman development company, the aim is to be able to book reserves for the field. StatoilHydro will not be able to carry corresponding reserves on its balance sheet, since it has not paid anything. But the booked reserves have no intrinsic value. Total and StatoilHydro will have the same cash flow from operation of the field. According to press reports, Total has thereby paid a substantial sum in order to improve its balance sheet – assuming that these reports are correct. The consequence is that the cash flow to Total shareholders will be weakened. StatoilHydro's shareholders are in the opposite position. Since the company has paid nothing in advance, net cash flow will be higher. But it must also live with lower booked reserves. However, it remains unclear how differences in payment could produce different rights for booking reserves – and how this relates to the relevant accounting rules. Will Total own reserves in the usual way as well as owning part of the development company? Will it acquire different rights from Statoil? Will it receive payment in a different way? The companies have not been allowed to make any further comments on the terms nor have the final frame conditions been established. Negotiations on actual participation in the Shtokman project took no less than 18 years. Lund told *Dagens Næringsliv* on 29 October 2007 that what has been concluded so far is a commercial frame agreement, and that he would provide more details in 2009.[11] Nor will the last word be said in 2009 – continuous renegotiation seems to be the guiding principle of Russia's petroleum administration. Moscow chief Hansen told the *Stavanger Aftenblad* daily that a bonus is to be paid in 2009 to participate in the project and

[9] http://www.dn.no/energi/article1211796.ece

[10] However, reports concerning the recognition of reserves in the field are conflicting. Oslo business daily *Dagens Næringsliv* reported on 10 January that StatoilHydro may be able to carry these reserves regardless.

[11] http://www.dn.no/energi/article1214983.ece

that this represents the point when the investment decision will be taken.[12] Experienced industry sources say that Total is a highly competent international player, and that the USD 800 million it has paid is probably not solely for the right to carry reserves but more of a regular signature bonus – and as such not particularly surprising. However, it is not entirely normal to begin conceptual studies for developing a field before the frame conditions have been settled. The impression one gets is that Total is in the driving seat for these studies. Is this a reflection of the fact that it has already paid a signature bonus, or the result of pure expertise considerations? Whatever the answer, it is a matter of concern if the two international participants in the field do not obtain the same incentive structure. During the award phase, the Russians demonstrated to the full that they are applying the principle of divide and rule. The question is whether they understand that running a licence in this way will be inappropriate once the award has been made. Have they grasped that constant renegotiation weakens incentives for the companies to make a long-term commitment to optimising value creation from the field?

4 Peregrino

Is StatoilHydro reserve-driven? Former Hydro's poor RRR is also making its mark on the merged company – proven reserves at 31 December 2007 were 6,010 million boe, compared with 6,101 million a year earlier. That represents a decline of 91 million boe. Reserves in 2007 grew by 542 million boe through revisions, extensions/expansions and new discoveries, compared with a growth of 383 million in 2006 from the same sources. The RRR was 86% in 2007, compared with 61% in 2006, while the average three-year replacement rate – including the effect of sales and acquisitions – was 81% at 31 December 2007 compared with 76% at the end of 2006.[13]

Does this put the company under pressure to obtain reserves quickly? Securing reserves through exploration is a time-consuming process and would not help alleviate the acute reserve problem. However, the company has an active exploration programme which is likely to contribute future additions to reserves. The short-term problem is that today's price level means reserves are very much a seller's market. By acquiring a 50% holding in Brazil's Peregrino heavy oil field in March 2008, the company would – according to certain analysts – be able to report a RRR of more than 100% for 2008. This also seems to have been a target for the company.[14] But little is certain when it comes to reserve bookings. Statoil learnt that previously when it had to write down its reserves in Ireland, as did Hydro when it wrote down

[12] http://web3.aftenbladet.no/innenriks/okonomi/article536237.ece

[13] http://www.statoilhydro.com/no/InvestorCentre/results/QuarterlyResults/Pages/FourthQuarter2007.aspx

[14] http://www.reuters.com/article/rbssEnergyNews/idUSL0940782820080109

the Spinnaker acquisition in the Gulf of Mexico. The same is applied this year. The RRR for 2008 ended at merely 34%.[15] Expected reserves, however, was probably fully replaced in 2008, if we use the company's own price forecasts. The booked reserves are seriously affected by the dramatic 70% reduction in the oil price, since reserves are booked according to the spot price.[16] By aggressive acquisitions StatoilHydro added large volumes of oil and gas to its portfolio, which will turn into booked reserves if prices go up again.

With reference to the Peregrino acquisition, oil commentator Arnt Even Bøe wrote in *Stavanger Aftenblad* on 5 March that StatoilHydro used to discover oil fields but is now buying them up – while prices are at a peak. According to Bøe, proper oil companies find their own reserves. However, he added that the acquisitions also contain a number of bright spots. According to StatoilHydro, experience off Norway with the Grane heavy oil field and drilling on Troll could provide a substantial increase in Peregrino's recovery factor. In addition, there come strategic considerations such as strengthening the company's core areas and securing the operatorship for the production phase. StatoilHydro was originally operator only for the development stage, with Anadarko due to take over once production began.

Many people would agree with Bøe that a long-term and sustainable oil company will primarily find oil through its own exploration efforts. This is where the greatest value creation occurs. Farming in and out of licences can be a favourable supplementary activity, but must then be counter-cyclical (buy cheap and sell expensive) rather than pro-cyclical.

To make money farming into licences at a time when oil prices are high, the company must be able to estimate reserves better than the seller or to develop and operate the field more efficiently. StatoilHydro has very extensive exploration operations both in Norway and abroad, and is likely to replace reserves by its own efforts over time. But the company faces a short-term problem with reserves. The question is then whether to bide one's time or make acquisitions. Virtually all the international oil companies are in the same boat after cutting back their exploration operations in the 1990s and also experiencing poor drilling results.

A good deal of information about the Peregrino acquisition is provided in a stock market announcement from StatoilHydro on 5 March 2008.[17] Expected reserves from this big heavy oil field are estimated at about 500 million barrels, excluding upsides. Production is scheduled to begin in 2010 and to provide StatoilHydro with additional output in the order of 100,000 barrels per day. The company already had a 50% holding in the field, which lies off Rio de Janeiro, and now becomes the sole licensee. StatoilHydro reported that the Peregrino project can cope with an oil price of less than USD 50 per barrel. At the same time, the purchase con-

[15] http://www.reuters.com/article/oilRpt/idUSLG57935420090217

[16] Also, reserve booking rules prevent StatoilHydro from booking the tar sand acquisition in Canada.

[17] This can be accessed at the company's website. See http://www.statoilhydro.com/en/NewsAndMedia/News/2008/Pages/Peregrino4March.aspx

tract has a clause worth recognising. StatoilHydro is paying NOK nine billion for the share of Peregrino and 25% of the deepwater Kaskida discovery in the Gulf of Mexico.[18] A possible additional compensation of up to NOK 1.5 billion may be paid for Peregrino if future oil prices are above predefined levels up to 2020. This shares the risk between buyer and seller. StatoilHydro has clearly hedged the downside through this agreement, but also appears to have ceded a substantial part of the upside.

In spite of higher oil prices, the average price paid for proven and probable reserves in the international oil industry was USD 4.67 per boe in 2007, down from USD 5.18 in 2006.[19] Higher oil prices have been more than offset by cost and tax increases. In an interview with *Dagens Næringsliv* on 4 March 2008, share analyst Gudmund Hille Isfeldt in DnBNor Markets estimated that StatoilHydro is paying USD 1.4 billion for Peregrino, plus an optional USD 300 million from 2010 to 2020 depending on oil price trends. USD 1.4 billion translates into a price of USD 5.60 per barrel, excluding the USD 300 million related to oil prices in the production period. Isfeldt added that the price per barrel becomes substantially lower when the upside in the reserves is taken into account.

Two aspects are of particular interest for a closer look.

1) After the acquisition, StatoilHydro will be the sole licensee.

Normal practice is for international oil companies to hold licences through joint ventures with each other. The advantages relate partly to operations and partly to risk sharing. More participants in a licence provide access to a wider range of expertise, and the companies can jointly arrive at optimum technical and commercial decisions. This also permits the sharing of project-specific risks, which can often be substantial – such as cost overruns and surprises related to the reservoir and production. It is accordingly unusual to be the sole licensee of a field of this size. The risk will quite simply be too large. An explanation for the acquisition could be that an increased holding provides greater potential for carrying reserves on the balance sheet. Another possible reason could be differences of opinion over the way the field should be developed. StatoilHydro has ambitions of achieving a higher recovery factor than would have been the case with the original plans, which also calls for much larger investment. The opportunity to bring in other licensees at a later date will nevertheless remain open, subject to approval by the authorities.

2) The payment for the licence transfer is a function of the future oil price.

The settlement for the licence share takes the form of a fixed amount plus a possible supplementary compensation of up to NOK 1.5 billion if future oil prices rise above predefined levels by 2010. Tying payments to future oil prices might be regarded as risk hedging at the project level. StatoilHydro reduces the amount it has to pay today

[18] The latter acquisition was subject to approval by the other partners in the license and has been turned down. This promising equity position is instead taken over by the partners.

[19] This emerges from a survey conducted by analysis company John S Herold and Standard Chartered Bank. See www.dn.nofor 11 March 2008.

in exchange for ceding part of the future upside in the project. However, risk hedging at project level would not be recommended on the basis of economic research. What means something to the owners of a company is its aggregate risk profile. Risk management should accordingly be based exclusively on assessments of the risk exposure of the company's overall portfolio. Since individual company projects will have risk profiles which cancel each other out to some extent, hedging need only be considered for part of the residual risk. If the company hedges at a lower level, such as a project, overall risk management could become excessive. This will lead in turn to sub-optimisation, and contribute in part to excessive transaction costs for hedging. It is otherwise also the case that investors who buy oil shares are precisely seeking to include oil price risk in their portfolio, and will react negatively if profits fail to grow sufficiently in line with rising oil prices. The possible unfortunate effects of the risk-sharing agreement on Peregrino – such as results failing to improve sufficiently as the price of oil increases – could however be reversed through the company's general risk management. One option would be transactions in the forward market. But this illustrates precisely the point that conducting risk management at two levels is pointless.

However, the agreement terms need not have anything to do with risk sharing. Licence farm-ins occur internationally where the settlement is conditional on specific outcomes (such as a specified level of oil prices). An optimum solution for two parties who take different views of the future could be to conclude such agreements.[20] If that is the case, it means that Anadarko has a more positive view of oil price trends than StatoilHydro.

The stock market announcement specified repeatedly that the acquisition was *strategic*. If this also means *expensive*, as experience would suggest, it could be appropriate to take a closer look at the agreed payment mechanism. In addition to the fixed settlement, StatoilHydro has given Anadarko an option conditional on the price of oil. Whether that is the intention, this helps camouflage the real breakeven price. Given the limited information provided, it is impossible to calculate the value of this option. At first glance, the acquisition looks cheaper than it actually is and people referred to a breakeven price of roughly USD 50 per barrel. The option payment must be added if the true breakeven price is to be identified. To achieve comparability with light oil projects – such as developments on the NCS – the spread between light and heavy oil must also be taken into account. The oil prices referred to in the press, Brent Blend and West Texas Intermediate (WTI), relate to light oil. At a press conference held after the acquisition, it was explained that the breakeven price of USD 50 cited by StatoilHydro for a heavy oil project referred to the Brent Blend reference crude, so comparability is maintained.

Heavy oil is priced considerably lower than light crudes, not least because of scarce refining capacity. Price trends for heavy crude could improve were capacity to be built up in the refining sector, but the development of a growing volume of

[20] Buyer and seller could achieve the same effect to some extent by taking positions in the forward oil market.

heavy oil reserves has prompted doubts among analysts about the progress of heavy crude prices. The spread between heavy and light oils at the time of the signing of the agreement was said to be USD 15–25 per barrel. Another specific project off Brazil operates with a spread of USD 23 per barrel. In other words, this amount must be deducted from quoted Brent Blend and WTI prices to find the heavy oil price.[21]

It was Hydro which acquired the first 50% of Peregrino (then called Chinook) for USD 350 million from Canada's EnCana in 2005. According to Isfeldt, StatoilHydro has paid USD 1.4 billion for the remaining 50% plus an option of USD 300 million from 2010 to 2020, depending on oil price developments.[22] We are talking here of a virtual quadrupling over three years. An increased recovery factor and higher oil price expectations play a big part, and StatoilHydro has upgraded the expected reserves during the development phase. But it appears that a good deal of strategic value may also have been assigned to the actual operatorship.

When Hydro acquired 50% of this licence in 2005, the recovery factor for this heavy oil field was estimated at 9%. With StatoilHydro's reservoir development plan, which utilises water injection and rock compaction, the estimated recovery factor rose to about 20%, and an even higher factor has been suggested later. That means estimated recoverable reserves have more than doubled.[23] When valuing this expansion, account must also be taken of the fact that increased reservoir utilisation has a substantial cost side. When assessing the value of reserves today compared with earlier valuations, it is important to determine whether the upgrades are based on new reservoir information. That appears to be only partly the case. The stock market announcement states that the potential supplementary resources are indicated by three-dimensional seismic surveys and have been partly proven by drilling a new well (3-PRG-0001-RSJ) in 2007. It also states that further appraisal wells will be needed to confirm remaining upsides in the south-western and southern extensions of the field.

The number of wells to be drilled and their spatial positioning – the well network – are of great significance for the recovery factor. But reservoir properties also mean a lot – the size of the residual oil saturation behind a water front, for instance. This can be difficult to estimate without a production history and measurements.

Historical experience in the oil industry indicates that oil companies overinvest when crude prices are high, and are therefore cautious about expressing high

[21] The discount on various oil grades depends on the supply and demand of a given grade and how many potential buyers can handle heavier oils. Where heavy crude is concerned, the discount will depend on how heavy it is, often expressed as degrees API, as well as on other factors such as its viscosity, how complex it is to refine, whether it could be blended with lighter oils to permit refining and so forth. Rather than a *single* spread, a whole range of prices exist. According to industry specialists, the Peregrino oil has an API around 14, with an expected sales price 25–30% lower than WTI.

[22] http://www.dn.no/energi/article1328359.ece

[23] See http://www.statoilhydro.com/en/NewsAndMedia/News/2007/Pages/PeregrinoOperatorship.aspx

breakeven prices for new projects. At the same time, they need additional reserves – which place them in a dilemma. StatoilHydro is in good company here, along with virtually all the major international oil companies. A possible solution is optimistic cost and reserve estimates. The latter incorporate various growth options in the form of improved recovery from the main reservoir and supplementary resources. StatoilHydro is far more optimistic for Peregrino in this respect than was Anadarko (and all other potential bidders), and this undoubtedly represents part of the basis for the transaction. On the other hand, the company is also highly competent in getting a lot out of fields. The recovery factor on the NCS is the highest in the world. However, sub-surface experts are doubtful about how much of this high NCS recovery should be attributed to advantageous natural conditions and how much to expertise. It has been claimed, for instance, that seawater injection in Ekofisk has not only hindered seabed subsidence but also affected the wettability of the chalk in a more water-wetting direction, and thereby improved recovery. Furthermore, it has transpired that a number of the large Norwegian sandstone reservoirs have a naturally mixed wettability, ensuring a very high recovery factor through water injection or natural water drive from the underlying aquifer.

5 Conclusion

International oil companies face problems replacing reserves through their own exploration and development activities. Reasons for this include a reduced exploration commitment in the 1990s, fewer large discoveries and reduced access to oil fields in regions with large resources.[24] The latter is often referred to as resource nationalisation, where a number of resource-rich countries and regions like Russia, Venezuela and the Middle East reduce the access for IOCs at high oil prices. Efforts are being made to compensate for replacement challenges through extensive purchasing of reserves. The danger is that such acquisitions are made at a high price. Sharply rising costs in the oil companies could represent a substantial challenge if oil prices were to decline significantly, which they have. A focus on reserves and volume could then be at the expense of profitability. This is a normal condition for the industry, which has historically overinvested when oil prices were high. However, many market players argue that the strong growth in demand for petroleum and the substantial problems faced in replacing reserves will result in a permanent upward shift in the oil price. A number of serious players went so far as to say that the price of crude could not fall below the very high 2008 level. They were wrong.

A number of producer countries – typically those with the biggest resources – are not prepared to cede ownership or control over their petroleum to foreign companies. This creates challenges for gross value creation, since control of resources is often closely related to incentives for maximising the value of reserves. It also

[24] On the other hand, access to gas is simpler.

limits opportunities for the international companies in these countries. However, there should be scope for establishing synthetic incentives which imitate to some extent those provided by normal licence terms. Both oil companies and producer countries stand to benefit from such a solution.

This article has reviewed two cases involving StatoilHydro: the Shtokman field off Russia and Brazil's Peregrino discovery. StatoilHydro has manoeuvred itself in a competent manner into key positions in Russia and Brazil, which are clearly among the most promising producer nations in the coming years. The company has established a close collaboration with Gazprom and Petrobras, and has acquired promising licences in these two countries. Since Shtokman and Peregrino will absorb big personnel and capital resources, however, they cannot simply be assessed on the basis of the strategic opportunities which they could open for further growth. They must also deliver in relation to StatoilHydro's on-going value creation. Analysts and the stock market have been lukewarm or negative to Shtokman and positive to Peregrino.

The problem with buying reserves in other countries is that one typically bids against companies with experience from the area (asymmetric information). One can then end up suffering the winner's curse – paying above the true value. StatoilHydro has had some experiences of that kind, e.g. the Spinnaker acquisition in the Gulf of Mexico. The opposite position prevailed in the Peregrino licence, however, in that StatoilHydro already had a 50% holding. This was perhaps part of the reason why the company wanted to become the sole licensee, which is unusual for such a large field. Ceding part of the upside to the seller through an option related to the sale is at the outset unfortunate from the shareholders' perspective. However, this of course depends on the alternative selling price without the option. Sometimes such options are crucial to trigger a sale. Moreover, StatoilHydro has acquired an operatorship where it can utilise its experience and expertise from similar developments. If the company succeeds in achieving high reservoir utilisation, as it has managed on the NCS, the investment will still provide an upside providing costs are kept under control. It could then also represent an important reference project for the company, which could make it easier to acquire other reserves. However, high reservoir utilisation calls for a lot of drilling, and rig rates are exceedingly high today. But it is possible that the substantial volume in the field could justify this, and rig rates are likely to go down. A high spread between light and heavy crude prices as well as special costs associated with recovering heavy oil could represent challenges for project economics. A good deal of environment-related uncertainty also attaches to heavy oil projects.

A way of overcoming the problem presented by asymmetric information when bidding for reserves would be to specialise in specific geographic areas and geological structures. That avoids having to bid constantly against companies who know more than oneself. Other considerations also favour a concentration, such as becoming familiar with regulations and their enforcement and establishing relations with the supplies industry. StatoilHydro has had a system of geographic core areas, but this does not always appear to have been effective in limiting the spread of activities.

Where Shtokman is concerned, StatoilHydro has entered into a contractor contract where its payment appears on paper to comprise a regulated maximum return for leasing production equipment over a 25-year period. This type of deal is more suitable for contractor companies. Its remuneration profile is not what investors in oil companies are looking for – namely, a cash flow which varies with production and gas prices. In addition to the long payback period in a country with substantial political risk, a substantial downside risk probably exists in relation to delays and overruns. Basically, there does not appear to be an upside which can compensate for the downside in the project. However, the commercial terms are still subject to negotiation, and efforts are being made to introduce synthetic incentives to the contract which will give StatoilHydro an upside related to the development of gas prices and the produced gas volumes. If such terms cannot be incorporated in a credible way (through having the contract refer to *international* gas prices, for instance), it is difficult to see why StatoilHydro should want to give final consent to the agreement in 2009. The Shtokman involvement will lay claim to many competent people in a period when expertise is in short supply, and will also call for very substantial capital outlays. These aspects must be balanced against corresponding upside opportunities. Compared with Total, StatoilHydro may have a strategic advantage in the final negotiation since it has not paid a signature bonus yet. Ultimately, however, both companies are dependent on the Russians sticking to their agreements. That does not appear to have been the case so far, but the Russians are in good company with other producer countries in this respect.

Russian authorities have so far had the advantage that oil companies, in their hunt for reserves, have been queuing up to develop fields in Russia. As experienced negotiators, they have also organised the playing field in such a way that the foreigners are pushing hardest for an agreement. However, negative experiences for foreign oil companies in Russia have shortened the queue to some extent. Moreover, plans and milestones for the Shtokman development now appear to have been established. It would not look so good for the Russians if StatoilHydro were to jump ship in 2009, which could give the latter a certain negotiating strength. This is the type of raw bargaining power which the Russians seem to understand. However, it remains unclear whether they fully comprehend that an agreement which provides StatoilHydro and Total with sufficient upside is necessary to harmonise their goals with those of the Russian authorities in order to achieve the largest possible value creation from the field. The willingness of the Russians to observe agreements is also questionable. As a result, it may be simpler in today's circumstances for the supplier companies to make money in Russia since they are paid on a continuous basis and can pull out should payment fail to be made. That will not be an option for StatoilHydro or Total once they have locked many billions of kroner into irreversible infrastructure investments.

While StatoilHydro can recognise booked reserves in Peregrino quickly, how far it will be able to do so with Shtokman remains an open question. Recognising reserves in the field will be possible in formal terms, and the Russian authorities would have nothing to lose from foreign companies doing so. Any barrier to recognising reserves would be raised by resource nationalism, but it is hard to believe that such

considerations would be stronger in Russia than in Iran. In any event, Shtokman cannot relieve reserve replacement challenges in the short term, since it is unlikely that the field can be booked as reserves for many years because technological, legal and financial conditions have yet to be clarified.

Acknowledgement I would express my thanks for rewarding conversations with and comments on the article from a number of key specialists in the oil industry and academia.

References

Arnott, R. (2004). Oil and gas reserves: communication with the financial sector. Paper on Sustainable Development Programme SDP BP 04/02, Oxford Institute of Energy Studies.

Kretzschmar, G.L., Misund, B., Hatherly, D. (2007). Market risk and oilfield ownership – refining SEC oil and gas disclosures. *Energy Policy, 35*(11), 5909–5917.

Lund, D. (1992). Petroleum taxation under uncertainty: contingent claims analysis with an application to Norway. *Energy Economics, 14*(1), 23–31.

Misund, B., Asche, F., Osmundsen, P. (2008). Industry upheaval and valuation: empirical evidence from the international oil and gas industry. *The International Journal of Accounting, 43*(4), 398–424.

Rajgopal, S. (1999). Early evidence on the informativeness of the SEC's market risk disclosures: the case of commodity price risk exposure of oil and gas producers. *The Accounting Review, 74*(3), 251–280.

Elastic Oil: A Primer on the Economics of Exploration and Production

Klaus Mohn

Abstract Predictions from the original geophysical approach to oil exploration and production suggest that oil production will develop according to a predetermined and inflexible bell-shaped trajectory, quite independent of variables relating to technological development, economics, and policy. Exploring the potential sources of elasticity in oil reserves and production, this paper offers a modification to the geophysical approach. Based on economic theory and modern empirical research the results suggest that both reserve-generation and production are indeed influenced by factors and forces of technology, economics, and government regulation.

1 Introduction

The sharp oil price increase over the last few years has increased the interest for security of energy supply in general, and for oil supply in particular. An important factor behind the oil price surge is strong economic growth in large parts of the world, but especially in newly industrialized economies like Brazil, Russia, India, and China. Another important factor relates to oil supply. So far, the response in oil supply to the latest price increase has been muted, partly due to financial pressures and enhanced capital discipline among international oil and gas companies (e.g., Osmundsen et al. 2007; Aune et al. 2007), but potentially also due to more fundamental factors relating to the non-renewable nature of fossil fuels. In this context, it is interesting to note that the most important petroleum provinces in the OECD area are faced with depletion (e.g., USA, Canada, United Kingdom, and Norway). International companies are therefore gradually shifting their attention and activities toward resource-rich countries in other parts of the world (e.g. Russia, Latin-America, Africa, and the Middle East).

K. Mohn
StatoilHydro (E&P Norway), 4036 Stavanger, Norway
and
Department of Economics and Business Administration, University of Stavanger, 4035 Stavanger, Norway
e-mail: kmohn@statoil.com

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_3,
© Springer-Verlag Berlin Heidelberg 2010

Ultimately, global oil reserves are bounded by nature, with physical limits both to availability and production growth. One of the early proponents of the geophysical approach to oil exploration and production was Hubbert (1962), who argues that cumulative production is the key predictor of the rate of production. According to this view, the geological knowledge which has been gained in a region is best described by cumulative production. As the region matures, cumulative production will also capture the inescapable destiny of depletion. And since production is determined by the level of reserves, reserve depletion will also cause an ultimate dampening of both investment rates and production. In consequence, petroleum production will develop according to a logistic growth function, yielding bell-shaped trajectories for exploration activity, reserve additions, and production. The sort of production profiles generated by the geophysical approach to petroleum exploration and production is illustrated in the right-hand panel of Fig. 1.[1]

Scale economies due to learning-by-doing (e.g., Quyen 1991; Krautkraemer 1998) will normally produce rapid growth in annual reserve additions from new discoveries in the early phase of development of a new oil province.[2] As the province matures, the average field size of new discoveries will tend down, and annual reserve

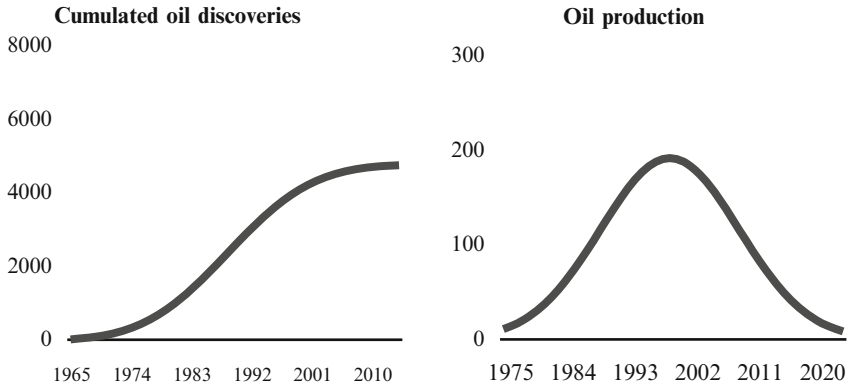


Fig. 1 The geophysical perspective on oil exploration and production. Source: Stylised example based on author's calculations

[1] The so-called Hubbert's peak was (quite successfully) applied to predict that US oil production would reach its maximum around 1970. The same concept has inspired the current debate of Peak Oil, with high-spirited discussions about when the world's oil production will peak.

[2] A popular analogy is found in the classic board game "Battleship". In the early phases of the game, with many ships on the board, expected rewards from bombing are high, with major learning effects involved whenever a new ship is hit. However, expected marginal gains, as well as learning effects, drop towards the end of the game, when the majority of ships have been sunk.

additions will diminish. This is illustrated in the left-hand panel of Fig. 1, whereby a bell-shaped curve for annual reserve additions gives rise to an s-shaped curve for cumulated volumes of discovered oil reserves.

As opposed to the scientists of geophysics and geology, economists like to think that oil production is governed by competitive companies' maximization of expected profits. Consequently, economists put special emphasis on the influence of unit costs of reserve-generation and production, market developments, and policy regulations. This does not imply that economists entirely neglect the geophysical aspects of oil exploration and production. Rather, the physical perspective represented by Hubbert's peak is regularly taken as a point of departure, and augmented with models and variables based on economic theory.

An obvious conundrum for the geophysical approach to oil production relates to the actual development of global reserves and production rates over the last decades. The fact is that proved global oil reserves have increased by 75% since the beginning of the 1980s. Annual rates of production have increased by nearly 40% over the same period, and remaining global reserve life has gone from 30 to 40 years over the last 25 year period.[3] The static approach implied by the Hubbert curve fails in explaining this development (e.g., Lynch 2002), and one important source of this shortfall relates to technological development (see also Watkins 2006). Improved technologies have improved the reserve and revenue potential for reserve and revenue-generation, not only from exploration activities (e.g., Managi et al. 2005), but also from new techniques for increased oil recovery in producing fields (e.g., Watkins 2002). At the same time, unit costs have been pushed down by technological progress. New solutions for exploration, development and production have implied a range of input-specific productivity gains, related to capital, labor, and energy. Economic models of oil exploration and production seek to embed these developments through appropriate mechanisms of technological progress, and through the incorporation of technology variables in empirical research.

When geology meets the market there are also prices involved. The supply from profit maximizing oil companies is determined by the equation between the oil price and marginal cost of production. Moreover, oil is an energy bearer that faces varying competition from other energy bearers, like coal, natural gas, and hydro-generated electricity. Finally, oil companies operate in variety of input markets, with direct exposure to varying costs of capital, labor, energy, materials, and other commodities. Consequently, oil investment and oil supply is likely to be influenced not only by the oil price, but also by a range of other energy prices, and potentially also by shifts and shocks in the input markets. To some degree, these mechanisms are also captured by economic models of oil supply.

Empirical studies of OPEC's role in the oil market have generally failed to establish firm evidence of stable cartel behavior. However, recent studies do acknowledge that some sort of collusion is taking place. The current discussion is more about which model of imperfect competition the oil price formation adheres to, and to the

[3] According to BP's Statistical Review of World Energy 2007.

stability issues of OPEC's market power (e.g., Smith 2005). Whatsoever, industry structure and macroeconomic management may have implications for incentives at the operational level. If the group of OPEC is seen as a dominant producer of an oil oligopoly, the strategic response to a lack of investment opportunities among their non-OPEC competitors does not necessarily imply an increase in OPEC investment (e.g., Aune et al. 2007). Moreover, oil investment and oil supply may not respond to high oil prices in countries dominated by national oil companies, as these companies may rather seek to stabilize government revenue than to maximize profits.

As the geophysical approach to oil production is focused entirely on the subsurface determinants of reserves and production, it also neglects the influence of government regulation. Governments play a role along the entire value chain of oil and gas companies. They control the access to exploration acreage, they approve any development project, they set the conditions for operations, and they design and impose systems of petroleum tax and government take. Moreover, governments also decide on how to manage petroleum resources, and not the least how to deal with resource revenues. It follows that economic models of oil production also require a role for government regulation and policies.

All in all, the geophysical approach to oil exploration and production is improved if the modeling framework is extended to include processes and variables concerning technology, markets, policy regulations, and market structure. Such an enhancement adds flexibility and elasticity to the geophysical approach. The result is a model that yields a better understanding of the interface between geology and economics, with improved predictions of both reserve-generation and production. This combined approach is illustrated in Fig. 2, where an interval of elasticity is added for both reserve additions and for oil production. The shaded areas indicate possible outcomes for oil exploration and production, depending on local and global factors of technology, prices, market structure, and policy regulation.

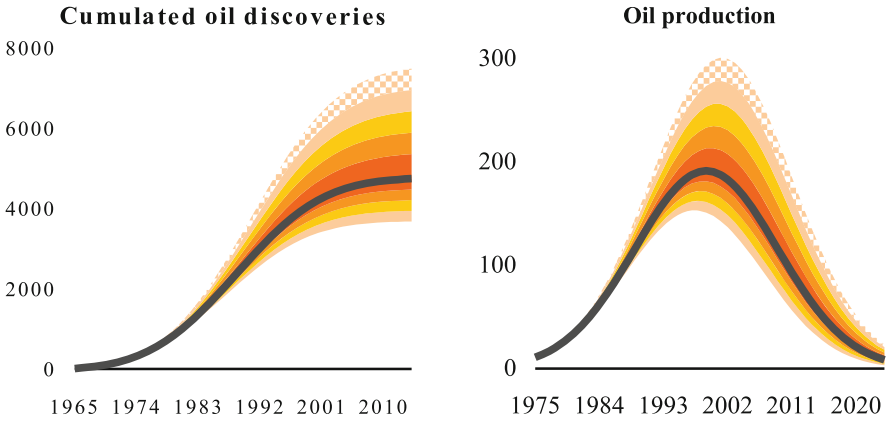


Fig. 2 The economic perspective on oil exploration and production. Source: Authors calculations

The remainder of this chapter is organized as follows. Section 2 provides a brief review of previous economic research on oil exploration and production, with a special emphasis on empirical models. To shed light on the economic approach to reserve-generation, Sect. 3 gives a retrospect on exploration activities on the Norwegian Continental Shelf (NCS). A couple of empirical models are demonstrated in Sect. 4, again based on data from the NCS. Concluding remarks are offered in Sect. 5.

2 Previous Research

As illustrated in Fig. 2, the economic perspective on oil exploration and production usually introduces a rightward bias in the logistic growth framework of the geophysical approach, as demonstrated in an empirical assessment of the ultimate resource recovery by Pesaran and Samiei (1995). This implies that over a period of time, resource additions tend to outpace the original geophysical estimates. In consequence, this also means that production rates will stay higher for longer than suggested by the simple Hubbert curve of Fig. 1.

One important source of this bias relates to technological progress. Technological progress can be addressed from two perspectives. On one hand, technological advances may exert a positive influence on the success rates in exploration and on the recovery rates of production. On the other hand, the dual approach is to view technological progress as a source of unit cost improvements. This would imply that technological advances induce an increase in yield per effort both in exploration and production. The accumulation of information and competence has the potential to improve the returns from exploration (e.g., Cleveland and Kaufmann 1997; Managi et al. 2005), as well as net revenues of production (e.g., Farzin 2001; Watkins 2002). In exploration, significant technological advances relate to the collection and interpretation of geological information, improved operational drilling efficiency, as well as new technologies for real-time monitoring and measurement of the well's downhole conditions. According to Forbes and Zampelli (2000), technological progress increased the success rate in US offshore exploration by 8.3% per year over the period 1986–1995. Similar advances in drilling technology are highly relevant also for production activities, as the investment in additional production wells become increasingly important when an oil field passes its peak, and embarks on the road towards depletion. Towards the tail-end phase of production, advanced reservoir management is combined with sophisticated drilling strategies to drain the reservoirs and to maximize resource recovery. Based on historical figures, Watkins (2002) finds that reserve appreciation over the lifetime of an average oil field amounts to some 20% for the United Kingdom, and close to 50% for Norway.

Another important deficiency in the original Hubbert model is its neglect of market mechanisms and price effects. Even though natural resources are bounded by nature, they are exploited by companies who adjust their behavior according to market developments and prices (e.g., Lynch 2002; Reynolds 2002). Empirical exploration models for the US oil and gas industry are surveyed by

Dahl and Duggan (1998), who conclude that acceptable models have been obtained for drilling efforts, with long-term oil price elasticities above one (see also Mohn and Osmundsen 2008; Ringlund et al. 2008). However, there is reason to believe that drilling efficiency is also influenced by the oil price, as risk propensity in company investment is affected by its financial flexibility (Reiss 1990; Iledare and Pulsipher 1999). Based on time series data for the Norwegian Continental Shelf, Mohn (2008) finds that reserve additions are indeed enhanced by an increase in the oil price, due to responses both in effort and efficiency of exploration. His explanation is that oil companies accept higher exploration risk in response to an oil price increase, implying lower success rates and higher expected discovery size. Since the beginning of the 1990s, a series of studies have also augmented the simple Hubbert approach to oil production with economic variables, most notably the price of oil (e.g., Kaufmann 1991; Cleveland and Kaufmann 1991; Pesaran and Samiei 1995; Moroney and Berg 1999; Kaufmann and Cleveland 2001). The research strategy of these studies has two stages. In the first stage, a reliable estimate is obtained for the Hubbert production curve. In the second stage, the deviation between observed production and the estimated Hubbert curve is modeled as a function of economic variables. All these studies show that economic variables are able to improve the quality of the original Hubbert model. However, the estimated oil price effects are modest, with elasticities of around 0.1 for the estimated production rates. The standard competitive model of supply has also been applied for empirical cross-country studies of oil supply (e.g., Watkins and Streifel 1998; Ramcharran 2002). In general, this class of models produces positive, but modest price elasticities for non-OPEC countries. On the other hand, the competitive model fails in providing a trustworthy description of OPEC supply.

The failure of competitive models in explaining OPEC supply behavior is simply a reflection of the imperfect competition in the global oil market. In 1960, OPEC was founded to unite the interests of petroleum policies across member states. Since then, OPEC oil ministers have met regularly to discuss prices and production quotas. In 2006, OPEC countries accounted for 42% of the world oil production and 75% of the world's proven oil reserves.[4] Empirical studies of OPEC's role in the oil market have generally failed to establish firm evidence of stable cartel behavior. However, recent studies do acknowledge that some sort of collusion is taking place. The current discussion is more about which model of imperfect competition the oil price formation adheres to, and to stability issues of OPEC's market power (e.g., Fattouh 2006).[5] A popular assumption for OPEC behavior is the target revenue hypothesis, which implies that production is regulated inversely with price to uphold a revenue level which is adequate for exogenous investment and consumption needs (e.g., Alhajji & Huettner 2000). The target revenue hypothesis imply that supply curves could be backward bending at high prices, which could again explain

[4] According to BP's Statistical Review of World Energy 2007.

[5] See Smith (2005) for a critical overview of empirical studies of OPEC behavior.

the muted investment response in OPEC countries to the current record oil price. However, as shown by Aune et al. (2007), net present value maximization combined with the exploitation of market power is also consistent with OPEC supply and oil price formation over the last years.

Finally, governments also exert an influence on reserve-generation and production in the oil industry. For profit-maximizing oil companies firms, profits are affected by tax systems and other forms of government take. Thus, incentives at the industry level may be affected by the regulatory system. In an econometric study of US exploration behavior, Iledare (1995) incorporates the tax system in his proxy for the marginal value of reserves. In exploration activities, governments also play an important role as the ultimate holders of exploration acreage. Access to exploration acreage is determined by licensing systems and policies, which therefore have to be incorporated in models of exploration activity and reserve-growth. Based on data from the NCS, Mohn and Osmundsen (2008) illustrate how exploration drilling is stimulated by awards of new exploration acreage, and Mohn (2008) also finds the size of average discoveries to be affected by licensing policies. Governments also play a role for the production phase of petroleum activity, with taxes and other systems of government take as the most notable transmission mechanism. As an example, a variable for pro-rationing of oil production in Texas prior to 1973 is included in Moroney and Berg's (1999) integrated model of oil supply. In general, tax systems have the potential of reducing investments and production growth (e.g., Boone 1998), distorting the optimal allocation of investments along the value chain,[6] and changing the distribution of capital for oil investment between countries. See Glomsrød and Osmundsen (2005) for a recent overview of these issues.

3 NCS Exploration and Production

The Norwegian Continental Shelf (NCS) is a relatively young oil and gas province. Its petroleum potential was ignited among geologists by the discovery of the Groningen gas field in the Netherlands in 1959. The first discovery on the NCS was made in 1969, and the Ekofisk field was put on stream 2 years later.[7] A number of discoveries were made in subsequent years (cf. Fig. 3), and, these laid the foundations for the evolution of a new and important industry in Norway, and a supplying region for US and European oil and gas markets. Today, 53 NCS fields contribute to the total Norwegian oil and gas production at 236 M standard cubic meters (scm)

[6] Capital requirement along the value chain include investments in exploration activities, field development, efforts to increase oil recovery, processing and transport facilities, and potentially also marketing activities.

[7] A non-commercial discovery (Balder) was actually made by Exxon (Esso) already in 1967. However, it took 30 years of technological development to mature this discovery into a profitable field development project based on subsea templates tied back to a floating production and storage vessel. The Balder field was put on stream in 1999 and is still producing (mid 2008).

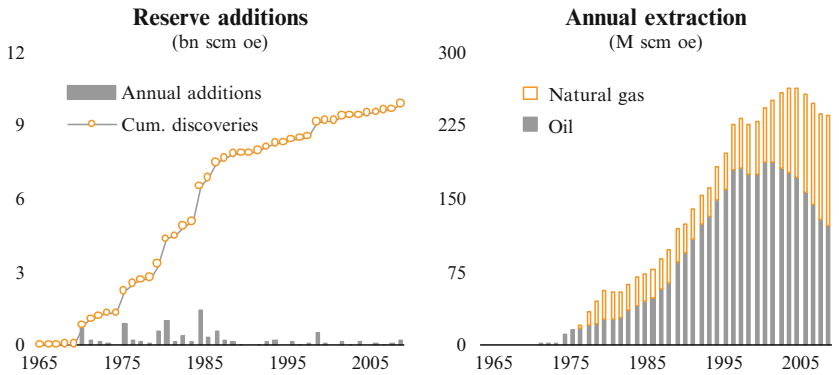


Fig. 3 NCS exploration and production. Source: Norwegian petroleum directorate

oil equivalents (oe), with a natural gas share of some 40% (2008). According to the Norwegian Petroleum Directorate (NPD), total oil production is now expected to continue its phase of gradual decline. On the other hand, gas production is seen to increase for another 5 years from today – to plateau levels of around 120 bn scm oe per year. For a thorough industry and policy overview of the NCS, see Ministry of Petroleum and Energy (2008).

Regulated gradualism has been a guiding principle for the development of the Norwegian oil and gas sector. The key regulatory instrument for exploration and production is the production license, which gives the exclusive right for exploration and production of oil and gas within a specified area, usually referred to as a block. Production licenses on the NCS are awarded through licensing rounds, and licensees retain ownership for the produced petroleum. A specific number of blocks are announced by the government, and the companies prepare applications based on published criteria. Based on submitted applications, the Ministry of Petroleum and Energy (MPE) decide on a partnership structure for each license, and an operator is appointed to take responsibility for the day-to-day activities under the terms of the license. Typically, a production license is awarded for an initial exploration period that can last up to 10 years. However, specified obligation regarding surveying and/or exploration drilling must be met during the license period. At completion of this kind of obligations, licensees generally retain up to half the area covered by the license for a specified period, in general 30 years.

After three decades of production, volume estimates from the Norwegian Petroleum Directorate (2007) indicate that 2/3 of the expected total physical oil and gas resources remain in the ground, nearly 40% of total resources are yet to be matured to proven reserves and 26% of total resources remain undiscovered. Exploration activity and results will be important to sustain production levels on the NCS over the longer term.

The first exploration well was struck in the North Sea in 1966, but it took 30 wells and 3 years before the breakthrough was made with the discovery of the Ekofisk field late in 1969. Since then, another 1,200 explorations and appraisal wells have been

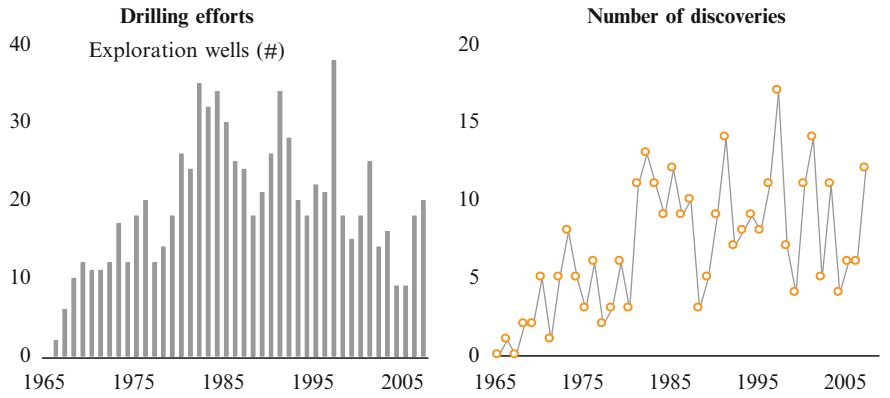


Fig. 4 Exploration efforts and number of discoveries. Source: Norwegian Petroleum Directorate and author's calculations

drilled, of which some 850 are classified as exploration wells (cf. Fig. 4). With more than 600 exploration wells, the North Sea represents approximately three-fourth of total cumulated exploration activity on the NCS. 160 exploration wells have been drilled in the Norwegian Sea, whereas only 63 exploration wells have been drilled in the under-explored Barents Sea. As illustrated in the right-hand panel of Fig. 4, the annual number of discoveries has largely hovered in the area between 5 and 15 over the last 20 years. However, we see a slight positive trend in the number of discoveries over time. To detect the sources and factors behind this development, the figures have to be decomposed even further.

A simple input measure to the exploration process is offered by exploration effort, or drilling activity, as illustrated in the left-hand panel of Fig. 4. A corresponding output measure is offered by reserve additions per exploration well (yield per effort). However, exploration output can be decomposed even further, as reserve additions per exploration well is the product of average discovery rate and average field size. The historical record for these two indicators is illustrated in Fig. 5. Over the 40-year period, the discovery rate has average 1/3, which is quite high by international standards. We also note that the volatility of the discovery rate was high in the early phase, which is an indication of high exploration risk due to inadequate information and poor experience. Over a period of time, however, discovery rates seem to have stabilized somewhat, and we also note a slight upward trend, which probably can be attributed to the accumulation of competence, experience, and technological progress.

The right-hand panel of Fig. 5 reports annual averages for the size of new discoveries. In their pursuit of maximum return, oil companies rank exploration prospects according to value potential, and target the structures with the highest potential first (Iledare 1995; Dahl and Duggan 1998). This is one explanation why the early phase of an oil province is usually dominated by large discoveries. However, government policies also played an important role for this development, as the early phase of

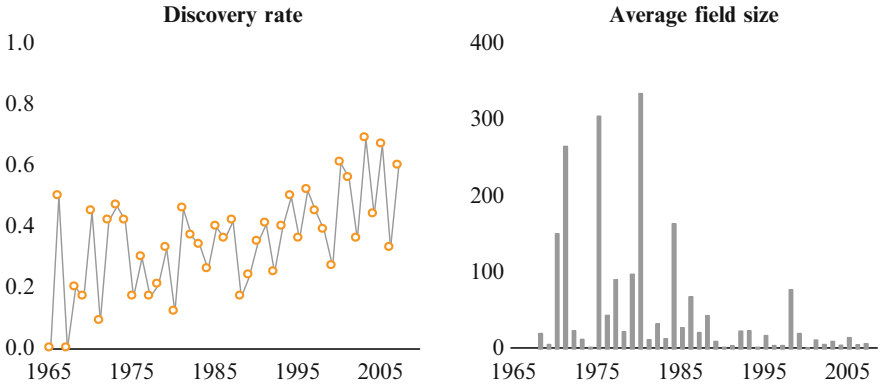


Fig. 5 Exploration success: discovery rates and average field size. Source: Norwegian petroleum directorate and author's calculations

the NCS history was characterized by regular licensing rounds, with a continuous supply of virgin exploration acreage with high potential. Finally, the high oil price level and an overconfident oil price outlook may also have induced oil companies to increase their exposure to overall exploration risk in the 1970s and early 1980s. This would also imply (lower expected discovery rates, and) a higher expected field size.

Over the last years, the NCS has gradually entered a more mature phase. Exploration efforts have been weak over the last years, and reserve additions from exploration have slowed to a trickle (cf. Fig. 3). Oil production has passed its peak, gas production is approaching its plateau, and there is no line of imminent new field developments. On the other hand, the record-high oil price provides a strong stimulus for investment to enhance recovery from the producing fields. To sustain investment and production over the longer term, the NCS will ultimately depend on new reserve additions from exploration. Both authorities and companies see a high potential for gas discoveries in the deepwater areas off Mid Norway, but so far, the establishment of a proven exploration play for this area is still pending (Norwegian Petroleum Directorate 2007).[8] Access to new exploration acreage with high potential in vulnerable waters off Northern Norway could be important to enhance the NCS reserve base. Due to environmental concerns, the issue of new awards in Northern Norway has developed into a highly controversial issue. However, an extrapolation of the Norwegian approach to petroleum management suggests that a political consensus will be reached, and that the industry will continue its gradual quest into Northern waters.

[8] An exploration play is a geographically bounded area where a combination of geological factors suggests that producible petroleum can be discovered. The three most important factors are (1) a reservoir rock where petroleum can be preserved, (2) a tight geological structure (a trap) that covers the reservoir rock, and (3) a mature source rock containing organic material that can be converted into petroleum (Norwegian Petroleum Directorate 2007).

4 A Simple Model of NCS Exploration

The reserve concept is one of the factors that distinguish non-renewable resource industries from other industries. Due to this defining characteristic, oil companies engage in extremely risky exploration activities to support and grow their base of oil and gas reserves, and to sustain production activity over the longer term. Among the oil companies, the set of exploration opportunities is subject to continuous evaluation and management based on a range of criteria relating to geology, technology, economic factors, and government policies. The result of this balancing act is a dynamic exploration strategy. Moreover, the implied portfolio of exploration drilling activities yields a certain average finding rate, a particular distribution of discovery size, and ultimately, a specific rate of gross reserve additions. Consequently, the data we observe for efforts and efficiency in oil exploration are formed by simultaneous decisions in each company. This simultaneity should be appreciated also in economic models of the exploration process. Drawing on Mohn (2008), an empirical modeling approach to exploration behavior will now be outlined, along with some results for time series data from the NCS.

The exploration process represents the traditional source for reserve additions. Based on sophisticated insight on the underground, exploration wells are directed at various layers that presumably hold oil and/or gas resources, according to different exploration plays. Exploration drilling may take place in virgin areas, where undiscovered deposits are potentially, and where the base of accumulated information and experience is correspondingly small. Alternatively, companies may focus their exploration in areas where fields have already been developed, with a significant base of competence and experience, and with access to well-developed infrastructure for processing and transport. Exploration in frontier areas represents higher risk than in mature areas. Within the companies, this risk is balanced against expected return in the management of the total exploration portfolio.

Based on standard principles from neoclassical theory for producer behavior, exploration activity may be represented by a standard production function, whereby inputs and technological progress are transformed into reserve additions. With profit maximization as the key behavioral assumption, such a model transforms into an optimal supply plan, where expected reserve additions depend on the oil price (P_t) and a set of state variables for geology (depletion; H_t), technology (Z_t), and government regulation (E_t). Having tested a range of alternatives, our preferred model includes cumulated exploration drilling activity as a proxy for depletion (H_t). Over the years, the collection of seismic data has grown exponentially on the NCS, reflecting the accelerating diffusion of increasingly advanced techniques for more efficient exploration activities. Accordingly, seismic surveying activity (Z_t) is included among our explanatory variables to capture technological progress. Finally, exploration efforts and efficiency is influenced by the availability of exploration acreage, which is subject to government regulation. Consequently, our model also includes the volume of open exploration acreage (E_t), which will be influenced by both licensing rounds ($\Delta E_t > 0$) and relinquishments and license expiration ($\Delta E_t < 0$).

We also bear in mind that reserve additions do not depend solely on efforts, but also on output. To this end, we apply a useful decomposition introduced by Fisher (1964), who demonstrated that annual reserve growth (R_t) can be seen as the product of exploration effort (D_t), the average discovery rate (S_t), and average discovery size (M_t). With explanatory variables grouped in the vector $\mathbf{X_t} = [P_t, H_t, Z_t, E_t]$, this yields for annual reserve additions:

$$R(\mathbf{X_t}) = D(\mathbf{X_t}) \cdot S(\mathbf{X_t}) \cdot M(\mathbf{X_t}) \,. \tag{1}$$

Equation (1) illustrates three sources of reserve additions, which all can be influenced by geology, technology, economics, and regulation. Consider the impact on reserve additions from an increase in the oil price. This will depend not only on how an oil price shock affects drilling activity (D_t), but also on its influence on the discovery rate (S_t) and average field size (M_t). The relation between these factors is again a result of the management of exploration portfolios within each oil company. To describe these mechanisms more precisely, define ε_P^R as the percentage increase in annual reserve additions caused by an oil price increase of one percent. Equation (1) now implies that this total elasticity can be represented by the sum of three partial elasticities:

$$\varepsilon_P^R = \varepsilon_P^D + \varepsilon_P^S + \varepsilon_P^M \,. \tag{2}$$

Thus, the impact of an oil price increase depends directly on how such an increase affects each of the three components of annual reserve generation. Corresponding elasticities apply for the other explanatory variables.

Applying simultaneous estimation techniques, Mohn (2008) now estimates the various elasticities implied by (1) and (2). Specifically, the empirical model has three endogenous variables (D_t, S_t, M_t), and is specified as a vector error-correction model, whereby changes in dependent variables are regressed on changes in explanatory variables, as well as the deviation from an underlying equilibrium relation between the model variables. Estimation is based on Full-Information Maximum Likelihood (Johansen 1995), as implemented in PcGive 10.

Key results for persistent elasticities are summarized in Fig. 6. The estimated long-term parameters illustrate the percentage impact on annual reserve additions from an increase in the explanatory variables of 1%. Moreover, Fig. 6 also illustrates how the combined elasticities of reserve growth with respect to explanatory variables may be decomposed, with partial attributions from drilling efforts (D), discovery rate (S), and average discovery size (M).

In terms of specific effects from explanatory variables, maturation and depletion (H_t) has a highly significant ($p = 0.00$)[9] dampening effect on annual reserve

[9] In testing of statistical hypotheses, the probability value (p-value) of a parameter estimate represents the likelihood of obtaining a result as extreme as the one obtained through our estimation, given that the null hypothesis is through. In our notation ($p = 0.00$), the implication is not that the p-value of this parameter estimate is actually 0, but that it fails to break zero at the two-digit cutoff level of measurement.

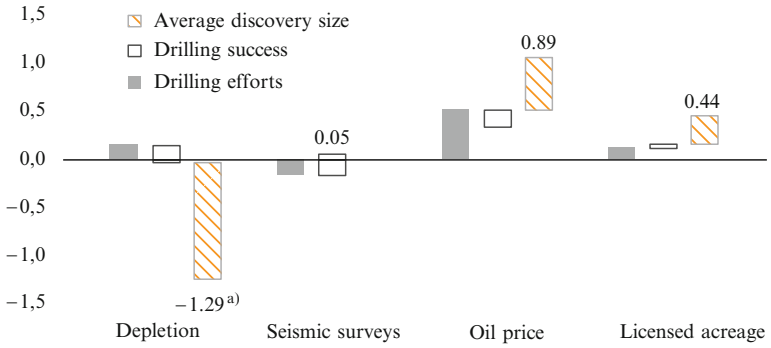


Fig. 6 Decomposed elasticities of reserve-generation. Estimated partial and total elasticities by explanatory variable (per cent). Key: * FIML estimates obtained with PcGive 10; [a] semi-elasticity: percentage change in reserve generation from abs. change in depletion indicator. Source: Mohn (2008)

additions, according to the estimation results. The main mechanism for this process is that the average field size falls over a period of time, which is also evident from Fig. 5. Seismic surveying activity, our proxy for technological development, has a mixed effect on the exploration process. The estimated total effect of this variable on reserve additions is small, and statistically insignificant ($p = 0.61$). However, the results do imply that seismic surveying activities contribute significantly to the increase over a period of time in discovery rates ($p = 0.00$), as indicated in Fig. 5.

Results for the oil price (P_t) illustrate the richness in economic effects from the proposed modeling framework, with statistically significant parameter for all the involved partial effects, as well as for the total effect ($p = 0.00$). Reserve additions are stimulated by an increase in the oil price, not only because drilling activities are spurred, but also because of positive effects on exploration efficiency – or yield per effort. Discovery rates are suppressed when the oil price increases, according to the econometric results. On the other hand, the estimated model establishes a positive and highly significant link between the oil price and average discovery size, an effect which dominates the estimated reduction in the discovery rate. This is a clear indication that oil companies adjust their portfolio of exploration activities according to changes in economic and financial conditions (Reiss 1990). In times of high oil prices, high cash-flows and high risk appetite, companies seem to tilt their exploration activities towards risky areas (frontier exploration), with relatively low discovery rates, and with high expected discovery size. When oil prices are low, cash flows are constrained, and the risk appetite is more modest, exploration strategies are typically more cautious. Consequently, exploration efforts are reduced, and focused in areas with higher discovery rates – and smaller expected field sizes (mature areas).[10]

[10] As opposed to frontier exploration areas, mature areas are typically characterized by proven exploration models, producing fields, well-developed infrastructure, transport facilities and market

Finally, the estimated model for the exploration process on the NCS also provides a significant role for government policies, as represented by access to exploration acreage. An increase in total licensed exploration acreage of 1%, will produce an increase in annual reserve additions by 0.44% ($p = 0.00$), according to the results. This effect has two sources. First, a modest increase in drilling activity is sustained when new acreage is offered. Second, new licensing rounds have a positive effect on average discovery size. With drilling efforts focusing on the most prospective available blocks at any time, it is natural that new licensing rounds will result in higher average discovery size.

The presented model leaves the impression that these variables relating to technology, economics and government regulation play a significant role for reserve additions on the NCS. Moreover, the outlined modeling approach provides a better representation of the complexity and sophistication of the exploration process than a simple geophysical approach. Consequently, the study by Mohn (2008) lends substantial support to the hypothesis that economic variables contribute to the explanation of oil exploration and production behavior. To illustrate this point more candidly, the presented model is re-estimated with the depletion indicator (H_t) as the only explanatory variable.

Table 1, reports the implied changes in estimated model quality, evaluated through the log-likelihood ratio (*LL*), as well as the Schwartz (*SC*), Hanna-Quinn (*HQ*), and Akaike information (*AIC*) criteria. These latter three criteria of model selection may be seen as goodness-of-fit measures for comparisons of different time series models based on the same data set. See Dornik and Hendry (2001) for theoretical background, technical detail and specific procedures for standard specification tests and model diagnostics in PCGive 10. At this point, we remind that an increase in the log-likelihood ratio (*LL*) is an indication of improved statistical power of the model, whereas a model improvement is generally associated with a reduction in the three other reported test criteria for model selection (*SC*; *HQ*; *AIC*). From Table 1 we clearly see that a disregard of economic variables yields a reduction in the log-likelihood, and increase in the other criteria of model selection. This confirms the preference for a combined model, and an appreciation of economic variables in the exploration process.

Table 1 The contribution of economic variables to overall model quality

Test statistics for model reduction				
	LL	SC	HQ	AIC
Presented model	26.50	0.42	−1.20	−0.47
Reduced model	−2.88	0.86	0.66	0.55

access. Moreover, exploration activities in these areas are usually directed at smaller satellite fields which can be tied back to already producing facilities of larger reservoirs (in decline), without the large investments involved by stand-alone field developments in new oil and gas regions (Norwegian Petroleum Directorate 2007).

5 A Simple Model of NCS Production

Previous research on oil and gas production suggests that economic variables may also improve the explanation of production activity. To test the impact of economic variables on extraction levels, Moroney and Berg (1999) propose and estimate a simple econometric model on data from the United States. Not surprisingly, they find that a combination of economic and geophysical variables provide an explanation which outperforms its alternatives, both in economic and statistical terms. Based on the framework of Moroney and Berg (1999) a model will now be outlined to perform a similar test on production data from the NCS (cf. Fig. 3).

As in the previous section, we first specify a model that contains both physical and economic variables. We then remove the economic variables, and compare the two model versions using both economic interpretation and statistical criteria of model selection. Consider a competitive firm that produces oil according to a well-behaved neo-classical production function $Q = F(\mathbf{L}, \mathbf{H})$,[11] where $\mathbf{L_t}$ represent a vector of variable inputs and $\mathbf{H_t}$ is a vector of state variables, including reserve variables, technological conditions and government policy. Maximization of profits (Π) now implies that the following restricted profit function can be derived:

$$\begin{aligned} \Pi = \Pi(P, \mathbf{W}, \mathbf{H}) &= \max_{Q,\mathbf{L}} \{P \cdot Q - \mathbf{W} \cdot \mathbf{L}\} \\ s.t.\ F(\mathbf{L}, \mathbf{H}) &\geq Q\ , \end{aligned} \tag{3}$$

where P is the price of oil, and $\mathbf{W}$ is the vector of input prices. Previous literature suggests that the role of traditional inputs is dominated by other factors in the process of oil and gas exploration and production (e.g., Dahl and Duggan 1998; Farzin 2001; Mohn 2008). The attention of this modeling exercise will therefore be focused on potential variables of the $\mathbf{H}$ vector, and the vectors of variable inputs and their prices are neglected for simplicity of exposition.[12] In this example, we are especially concerned with the role of economic variables (P) as opposed to geological variables. Consequently, the $\mathbf{H}$ vector of this sketchy application will therefore be earmarked for variables of reserve development and depletion.

In our approximation of an empirical specification for oil production, we now assume a multiplicative form for the restricted profit function:

$$\Pi = \Pi(P, H) = \tilde{A} P^{\tilde{\alpha}} \exp[\beta H + \gamma H^2]. \tag{4}$$

[11] With a long-term perspective on the production process, all inputs may be seen as variable. Consequently, the capital stock can be included in both the L and the H vector, depending on the horizon of the decisions in question.

[12] To test the validity of this assumption, a variety of interest rate and labor cost variables were included in preliminary estimations. However, plausible and robust estimates could not be established for any of their coefficients.

Observe that a squared term is included for the depletion mechanism, to allow for potential non-linearities in the process of resource exhaustion. Hotelling's lemma now allows the derivation of optimal oil supply directly from (4). Partial differentiation with respect to the oil price now yields:

$$\frac{\partial \Pi}{\partial P} = AP^{\alpha} \exp[\beta H + \gamma H^2], \tag{5}$$

where $A = \tilde{\alpha}\tilde{A}$ and $\alpha = \tilde{\alpha} - 1$, H is a depletion indicator, proxied by accumulated production, and α, β, and γ are the coefficients to be estimated.

Introducing small-caps for natural logs, as well as a time index t, we now specify the econometric model as a simplified error-correction representation of (5):[13]

$$\Delta q_t = \lambda q_{t-1} + b_0 p_{t-1} + b_1 H_{t-1} + b_2 H_{t-1}^2 + u_t. \tag{6}$$

The underlying structural parameters of (3) can be calculated directly from the estimated parameters of (6).[14] The lag structure of (4) implies a gradual adjustment to oil price changes which is consistent with adaptive price expectations (cf. Farzin 2001). A simple form of expectations formation is therefore encompassed by the error-correction specification. This specification also removes problems due to non-stationarity in the model variables, and secures dynamic balance among variables in the econometric equation. Equation (6) may therefore be estimated by ordinary least squares.

Based on annual time series data over the period 1972–2004, we obtain:

$$\Delta \hat{q}_t = \underset{(0.00)}{-0.41} q_{t-1} + \underset{(0.00)}{0.38} p_{t-1} + \underset{(0.00)}{0.95} H_{t-1} \underset{(0.00)}{-0.19} H_{t-1}^2 . \tag{7}$$

The estimated model is well-behaved, and meets the requirements implied by standard specification tests. All parameters are highly significant in statistical terms, as indicated by the p-values (in brackets). The lagged production level exerts a negative influence on production growth, according to the estimated model, implying a slowdown in production as long as production increases. Observe also that our depletion indicator takes a positive coefficient, suggesting that production growth is actually stimulated by cumulated production. However, this stimulation is modified by the negative and highly significant coefficient on the squared depletion term. In sum, the standard properties of the geophysical approach seem to be fairly well

[13] The error-correction specification would normally also include changes in model variables. However, these proved insignificant in preliminary estimations, and are therefore left out for simplicity of exposition. The constant term is also removed for the same reason.

[14] Letting all changes approach zero, (4) can be solved for q_t to obtain $\alpha = -b_0/\lambda$, $\beta = -b_1/\lambda$, $\gamma = -b_2/\lambda$ (Bårdsen 1989).

captured by (7). Observe also that the oil price takes a positive coefficient. This indicates that production levels on the NCS are significantly influenced by this key economic parameter. To study the contribution of economic variables to our explanation of NCS production growth, we now reestimate the model, leaving out the price of oil. This yields:

$$\Delta\hat{q}_t = \underset{(0.30)}{0.05} q_{t-1} - \underset{(0.61)}{0.17} H_{t-1} + \underset{(0.33)}{0.03} H_{t-1}^2 \,. \tag{8}$$

The first impression is already that (8) provides quite a miserable explanation of production activity compared to (7). All parameters approach zero, they change signs, and none of them are significant in statistical terms.

The standard criteria of comparison is the squared multiple correlation coefficient R^2. However, as the constant term was removed from our preferred model based on statistical inference, R^2 is no longer well-defined.[15] The equation standard error provides a better statistic for model comparison, as this measure is adjusted for degrees of freedom. For (7), we obtain an equation standard error of 0.25, whereas the corresponding estimate for (8) is 0.45. This suggests a clear preference for the model represented by (7). To conclude even more rigorously, Table 2 compares the common battery of specification tests for statistical performance. As for the exploration models in the previous section, Table 2 reports changes in estimated model quality, evaluated through the log-likelihood (*LL*), as well as the Schwartz (*SC*), Hanna-Quinn (*HQ*), and Akaike information (*AIC*) criteria. These indicators are the same as applied for the exploration model above. See Dornik and Hendry (2001) for details on their properties. We remind that an increase in the log-likelihood ratio (*LL*) is an indication of improved statistical power of the model, whereas a model improvement is generally associated with a reduction in the three other reported test criteria for model selection (*SC*; *HQ*; *AIC*). Again, we see that a disregard of economic variables yields a substantial reduction in the log-likelihood, and increase in the other criteria of model selection. Consequently, we should prefer a combined model. In summary, the value and importance of economic variables in models of petroleum activity is further corroborated.

Table 2 The contribution of economic variables to the production model

Test statistics for model reduction				
	LL	SC	HQ	AIC
Model with oil price	0.77	0.38	0.26	0.20
Model without oil price	−19.21	1.48	1.39	1.35

[15] R^2 also has a range of weaknesses with respect to model evaluation. The inclusion of additional variables will never reduce the value of R^2, and it may improve even if nonsense variables are adjoined. Moreover, R^2 also depends on the choice of transformation of the dependent variable (for example, Δy versus y). R^2 may therefore be misleading for model evaluation purposes.

6 Concluding Remarks

Empirical research in petroleum economics has demonstrated again and again that predictions based on the geophysical approach to oil exploration and production gives a poor representation of actual development of the last 50 years or so. The typical pattern for individual fields, regions and provinces is that exploration activities uncover far more oil reserves than is thought possible in the initial estimates. Moreover, the accumulation of technology, experience and competence also makes it possible to recover more oil from each producing field than implied by the static traditional geophysical approach to oil extraction.

My two applications are based on data from the Norwegian Continental Shelf, an oil province whose development is characterized by gradualism and government regulation. New insights into the complex process of oil exploration are obtained through the combination of physical and economic variables in an integrated dynamic time series model. As an example, the results imply that additions to the reserve base are affected by the oil price, not only because drilling efforts are spurred when the oil price increases, but also because the output of the drilling process is influenced by prices, cash-flows and adjustments to the exploration portfolio in each company. The presented model of NCS oil exploration suggests that companies increase their exposure to exploration risk when the oil price goes up, yielding lower discovery rates, and higher average discovery size. On the other hand, a reduction in the oil price makes oil companies more cautious. A low oil price makes them focus exploration activities in less risky (mature) areas. The result is higher discovery rates, and smaller discoveries. In the same model, reserve depletion exerts a dampening effect on reserve growth, partly offset by the positive impact of seismic surveys on discovery rates. Through the design and execution of licensing rounds and awards of new exploration acreage, the expected reserve and production potential on the NCS is also affected by government policy. Awards of new exploration acreage give a stimulus to reserve additions due both to enhanced drilling efforts and improved drilling efficiency, according to the results.

A simple econometric example for NCS oil production also suggests that economic variables play a significant role in the explanation of production levels. The preferred econometric model of oil supply includes a positive and highly significant parameter for the real oil price, indicating an own-price elasticity of oil supply above 0.9, which is high by comparable standards. As for the exploration model, we also find the estimated model of production to deteriorate when this simple economic parameter is left out of the equation. A variety of specification tests and standard statistics of model fit and clearly suggest that simple geophysical models are outperformed by models which also include economic variables.

In summary, modern economic research has established a firm role for economic variables in models of oil exploration and production. The importance of technology, economics and policies to supplement the geophysical aspects of oil production is also supported by the two examples of/in this chapter.

References

Alhajji, A. F., & Huettner, D. (2000). The target revenue model and the world oil market: empirical evidence from 1974 to 1994. *The Energy Journal, 21*(2), 121–144.

Aune, F. R., Mohn, K., Osmundsen, P., & Rosendahl, K. E. (2007). *Industry restructuring, OPEC response – and oil price formation*. Discussion Paper 511, Research Department of Statistics, Norway.

Boone, J. P. (1998). The effect of the corporate alternative minimum tax on investment in oil and gas exploration and development. *Journal of Energy Finance and Development, 3*(2), 101–128.

Bårdsen, G. (1989). Estimation of long-run coefficients in error-correction models. *Oxford Bulletin of Economics and Statistics, 51*, 345–350.

Cleveland, C. J., & Kaufmann, R. K. (1991). Forecasting ultimate oil recovery and its rate of production: incorporating economic factors into the models of M. King Hubbert. *The Energy Journal, 12*(2), 17–46.

Cleveland, C. J., & Kaufmann, R. K. (1997). Natural gas in the U.S.: how far can technology stretch the resource base? *The Energy Journal, 18*(2), 89–107.

Dahl, C., & Duggan, T. E. (1998). Survey of price elasticities from economic exploration models of US oil and gas supply. *Journal of Energy Finance and Development, 3*(2), 129–169.

Dornik, J. A., & Hendry, D. F. (2001). *Modelling dynamic systems using PCGive 10*. London: TCL.

Farzin, Y. H. (2001). The impact of oil prices on additions to US proven reserves. *Resource and Energy Economics, 23*, 271–291.

Fattouh, B. (2006). *OPEC pricing power*. Working Paper 31, Oxford Institute for Energy Studies (http://www.oxfordenergy.org/pdfs/WPM31.pdf).

Fisher, F. M. (1964). *Supply and costs in the U.S. oil and gas industry: two econometric studies.* Baltimore: Johns Hopkins Press.

Forbes, K. F., & Zampelli, E. M. (2000). Technology and the exploratory success rate in the U.S. offshore. *The Energy Journal, 21*(1), 109–120.

Forbes, K. F., & Zampelli, E. M. (2002). Technology and the exploratory success rate in the U.S. onshore. *Quarterly Journal of Economics and Finance, 42*(2), 319–334.

Glomsrød, S., & Osmundsen, P. (2005). *Petroleum industry regulation within stable states*. Aldershot: Ashgate.

Hubbert, M. K. (1962). *Energy resources: a report to the committee on natural resources of the National Academy of Sciences*. Washington: National Research Council. Publication 1000-D. National Academy of Sciences-National Research Council.

Iledare, O. O. (1995). Simulating the effect and policy incentives on natural gas drilling and gross reserve additions. *Resource and Energy Economics 17*, 261–279.

Iledare, O. O., & Pulsipher, A. (1999). Sources of change in petroleum drilling productivity in onshore Lousiana in the US, 1977–1994. *Energy Economics, 21*, 261–271.

Johansen, S. (1995). Likelihood-based inference in cointegrated vector auto-regressive models. Oxford: Oxford University Press.

Kaufmann, R. K. (1991). Oil production in the lower 48 states: reconciling curve fitting and econometric models. *Resources and Energy, 13*, 111–127.

Kaufmann, R. K., & Cleveland, C. J. (2001). Oil production in the lower 48 states: economic, geological, and institutional determinants. *The Energy Journal, 22*(1), 27–49.

Krautkraemer, J. A. (1998). Nonrenewable resource scarcity. *Journal of Economic Literature, 36*, 2065–2107.

Lynch, M. (2002). Forecasting oil supply: theory and practice. *The Quarterly Review of Economic and Finance, 42*, 373–389.

Managi, S., Opaluch, J. J., Jin, D., & Grigalunas, T. A. (2005). Technological change and petroleum exploration in the Gulf of Mexico. *Energy Policy, 33*(5), 619–632.

Ministry of Petroleum and Energy. (2008). Facts 2008: The Norwegian petroleum sector (http://www.petrofacts.no).

Mohn, K., & Osmundsen, P. (2008). Exploration economics in a regulated petroleum province: the case of the Norwegian Continental Shelf. *Energy Economics, 30*(2), 303–320.

Mohn, K. (2008). Efforts and efficiency in oil exploration: a vector error-correction approach. *The Energy Journal, 29*(4), 53–78.

Moroney, J. R., & Berg, M. D. (1999). An integrated model of oil production. *The Energy Journal, 20*(1), 105–124.

Norwegian Petroleum Directorate. (2007). The petroleum resources on the NCS 2007 (http://www.npd.no/English/Frontpage.htm).

Osmundsen, P., Mohn, K., Misund, B., & Asche, F. (2007). Is oil supply choked by financial markets? *Energy Policy 35*, 467–474.

Pesaran, M. H., & Samiei, H. (1995). Forecasting ultimate resource recovery. *International Journal of Forecasting, 11*, 543–555.

Quyen, N. V. (1991). Exhaustible resources: a theory of exploration. *Review of Economic Studies, 58*, 777–789.

Ramcharran, H. (2002). Oil production responses to price changes: an empirical application of the competitive model to OPEC and non-OPEC countries. *Energy Economics, 24*, 97–106.

Reiss, P. C. (1990). Economic and financial determinants of oil and gas exploration. In M. K. Hubbard, & R. Glenn (Eds.), *Asymmetric information, corporate finance and investment*. Chicago: University of Chicago.

Reynolds, D. B. (2002). Using non-time series to determine supply elasticity: how far do prices change the Hubbert curve? *OPEC Review, 26*(2), 147–167.

Ringlund, G. B., Rosendahl, K. E., & Skjerpen, T. (2008). Does oilrig activity react to oil price changes? – An empirical investigation. *Energy Economics, 30*(2), 371–396.

Smith, J. L. (2005). Inscrutable OPEC? Behavioral tests of the cartel hypothesis. *Energy Journal, 26*(1), 51–82.

Watkins, G. C. (2002). Characteristics of North Sea oil reserve appreciation. *The Quarterly Review of Economics and Finance, 22*, 335–372.

Watkins, G. C. (2006). Oil scarcity: what have the past three decades revealed? *Energy Policy, 34*, 508–514.

Watkins, C. J., & Streifel, S. S. (1998). World crude oil supply: evidence from estimating supply functions by country. *Journal of Energy Finance and Development, 3*(1), 23–48.

Applied Mathematical Programming in Norwegian Petroleum Field and Pipeline Development: Some Highlights from the Last 30 Years

Bjørn Nygreen and Kjetil Haugen

Abstract This chapter discusses various attempts to apply mathematical programming tools and techniques in field development planning for the Norwegian continental shelf. The paper has a form of a (non-complete) survey, with the aim of discussing and presenting various attempts, both within deterministic and stochastic modelling.

1 Introduction

As indicated by Fig. 1[1], Norwegian oil production reached its historic peak level at around the year 2000. According to the predictions of Fig. 1, the production will reach insignificant amounts in 25 years or less from now. As Fig. 1 also indicates that the history of Norwegian oil production is relatively a recent one – production started less than 40 years ago – 15 June 1971. The first field discovered on the Norwegian continental shelf, The Ekofisk field, was discovered by Phillips Petroleum Company in 1969. Amazingly, it is still (2006) the largest producing unit among 51 producing units totally.

During this relatively short time, a significant amount of mathematical programming models and tools have been applied. Surely, classical applications in reservoir modelling and in refinery planning are among such applications, but this chapter will focus on the use of mathematical programming tools in investment/development planning for fields and pipelines. This is perhaps somewhat special for Norway, where such models have had a widespread use among companies and regulating

[1] see http://www.norge.se/business/oil/oilproductionno.htm

B. Nygreen (✉)
Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway
e-mail: bjorn.nygreen@iot.ntnu.no

K. Haugen
Molde University College, Box 2110, 6402 Molde, Norway
e-mail: kjetil.haugen@himolde.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_4,
© Springer-Verlag Berlin Heidelberg 2010

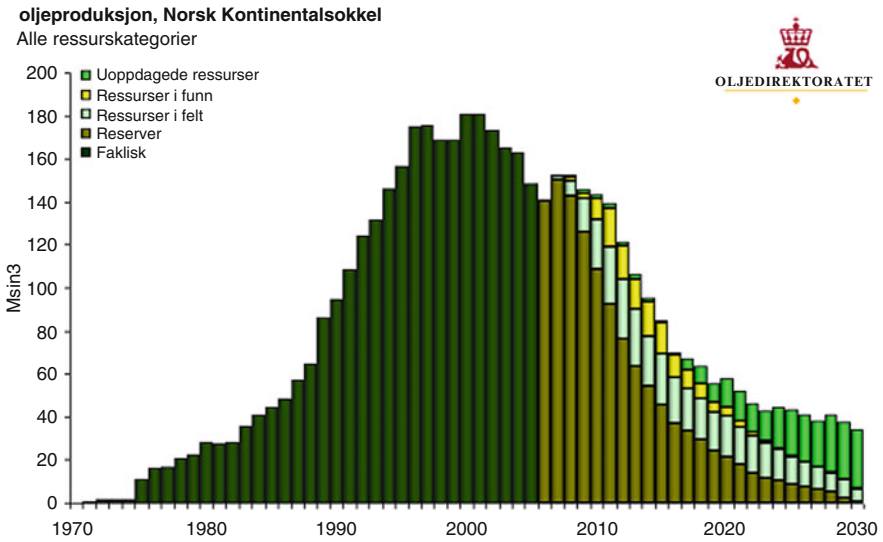


Fig. 1 Norwegian oil production over time

authorities. One important reason for this use may be the Norwegian petroleum legislation with a relatively large emphasis on governmental regulation of the whole industry. Still, research literature is relatively sparse on descriptions and reports of the above alleged model use, with some noteworthy exceptions. General examples (not necessarily coupled to the Norwegian case) may be found in Aronofsky and Williams (1962), Beale (1983), McFarland et al. (1984), Sullivan (1988), and Bodington and Baker (1990), while some other authors who focus explicitly on Norwegian field development planning are Haugland et al. (1988), Aboudi et al. (1989), Haugen (1991, 1996), and Nygreen et al. (1998).

In the next paragraph, we will give some short historic remarks; while in subsequent paragraphs, we will focus on modelling content on both deterministic and stochastic mathematical programming modelling in the Norwegian petroleum sector.

2 Some Historic Remarks

After the discovery of the previously mentioned Ekofisk field in 1969, several research groups in Norway have applied mathematical programming to solve different types of planning problems related to petroleum production. These groups can be divided into three types:

1. Groups within petroleum companies operating on the Norwegian Continental Shelf
2. Groups in applied research organizations
3. Groups in Norwegian universities and colleges including their students. These groups have also collaborated

Of the applied research groups, we want to mention two. The first is Christian Michelsen Institute (CMI) in Bergen, where Professor Kurt Jörnsten worked with planning problems in connection with petroleum in the 1980s. The department of CMI that Professor Jörnsten worked for became a limited company named Christian Michelsen Research in 1992. Part of his work at CMI is published in Aboudi et al. (1989). The second group is SINTEF in Trondheim, where the authors of this paper worked full or part-time in the 1980s also with planning problems for the petroleum industry. We participated in making a model not very different from the mentioned CMI model. Our model was published in Nygreen et al. (1998), after it had been in professional use for 15 years.

After the discovery of petroleum in the Norwegian continental shelf, several universities and colleges in Norway started to develop courses with greater emphasis on problems relevant for the petroleum industry. This was also the case for the Department for managerial economics and operations research at the Norwegian Institute of Technology (now a part of Norwegian University of Science and Technology) where both the authors of this paper have worked full or part-time most of this period.

During this period, several students have been offered master thesis work for companies engaged in the Norwegian petroleum business. We will not mention these students by name, but we want to mention that all of the 'pioneer students' using optimization models on problems in the North Sea at our department, studied applied physics with specialization in operations research.

In the forthcoming paragraphs, we will try to sum up some of the basic principles used in the above mentioned models emphasizing structural rather than operative modelling.

3 Main Deterministic Model Assumptions

The decision problems analysed in the early 1980s were deterministic problems where the discrete decisions – which fields to develop and when – were recognised as being the most important.

A typical model would contain a given number of projects which could be started in any one of several years. Some of these projects could also be forced to start within the given planning horizon. Some projects produced oil and/or gas, while other projects had capacities for processing and/or transportation of the products. When a project was decided to be started in a given period (model-wise), this would typically determine the amount of oil and gas produced by the project in all future years. All resources needed by the project in all future years, were also determined by deciding the start year.

Later these models were changed such that production was allowed to vary more flexibly with the start year. Hence, a certain element (within simple physical boundaries) of variable production opportunities was introduced.

4 Some Representative Deterministic Modelling

In order to indicate the basic modelling principles of the models discussed above, we describe a simplified version of the model published in Nygreen et al. (1998). We use lower case letters to represent subscripts and variables, and capital letters to represent constants and parts of the constant names written as literal subscripts.

Only essential parts of the implemented model will be described here. We have chosen not to comment on whether the constants are given directly or calculated from several constants that are not defined.

For simplicity, many variables and constraints will be defined for all the possible combinations of their subscripts, even if they are only used for some of the possible combinations. This means that the exclusion tests for variables and constraints are omitted. Sets describing which subscripts to sum over in various constraints are also omitted.

4.1 Start of Projects

If project j starts in period u, then variable x_{ju} is set equal to one. A project that does not start at all has its non-start variable, y_j, set equal to one. The user can set individual start periods for each project and also force some projects to start. This is not shown explicitly in the formulation. All projects may at most start once. This is modelled by declaring that all variables for project j in constraint type (1) belong to a special ordered set of type 1.

$$\sum_u x_{ju} + y_j = 1, \forall j \tag{1}$$

Some projects are dependent on the start of other projects. This is modelled through precedence constraints. In each such constraint, the dependent project is named the successor and the other project is named predecessor. If there is no bound on the difference between the start-time for such projects, we can write the constraints as in (2). We use d as subscript for the precedence constraints. In the following constraints, j is always the successor project, while i is the predecessor in constraint d.

$$y_j - y_i \geq 0, \forall d \tag{2}$$

4.2 Alternatives

The user can define several different ways to develop a given petroleum field by defining each possible development as a project and all these projects as an

alternative. An alternative is a set of projects where at most one is allowed to start. The number of projects in alternative a is N_{ALTa}.

$$\sum_j y_j \geq N_{ALTa} - 1, \forall a \tag{3}$$

The summation over j above is only for projects that belong to alternative a.

4.3 Production Decisions

In the first models, the production from each project was forced to follow the production profile given by user input from the decided start year. If that made it difficult to get the production to fit the capacities of the pipes and the markets, one had to insert some flexibility by allowing the production to deviate from the given profiles. Sullivan (1988) interpolates between two different developments of a field. Beale (1983) used variables for pressure both in the reservoir and in the pipes instead of profiles.

In a simple way, we may say that we chose the possibility of saving some product to a later period – at the same time avoid exceeding the maximal production in the profile determined by the user input.

The user specified the following data for production projects, directly or indirectly; T_{Pjs} = maximal total production of product s from project j, P_{Rjstu} = maximal production of product s from project j in period t if the project was started in period u, R_{ALjs} = factor to multiply P_{Rjstu} to get the minimal production of s in year t.

With p_{jst} as the variable for the produced amount of product s from project j in year t, and q_{jst} as the corresponding cumulative production we can write (a simplified and too restrictive version of) the production constraints as:

$$p_{jst} - \sum_u P_{Rjstu} x_{ju} \leq 0, \forall j, s, t \tag{4}$$

$$p_{jst} - \sum_u R_{ALjs} P_{Rjstu} x_{ju} \geq 0, \forall j, s, t \tag{5}$$

$$p_{jst} + q_{js(t-1)} - q_{jst} = 0, \forall j, s, t \tag{6}$$

$$q_{jst} \leq T_{Pjs} \forall j, s, t \tag{7}$$

Usually we cannot vary production of different products independent of each other. The constraints controlling this have been omitted here.

4.4 Global Constraints

In the early days of the Norwegian petroleum production, Norwegian politicians discussed possible bounds on both the total aggregated petroleum production each year as well as on total yearly investments on the Norwegian shelf.

For each project j, the user has given data, R_{jbtu}, for the usage of resource b in period t given that the project starts in period u. If project j is a production project, we also need a constant, R_{Pjsb}, which gives the amount of resource b for each unit of product s produced from the project. We define a variable, r_{bt}, which is equal to the total amount of resource b in period t.

$$\sum_j \sum_u R_{jbtu} x_{ju} + \sum_j \sum_s R_{Pjsb} p_{jst} - r_{bt} = 0, \forall b, t \tag{8}$$

In the case with aggregated production, we let b also represent the aggregate. Then R_{jbtu} is zero and R_{Pjsb} is aggregation factors for product s. Then let R_{MXbt} and R_{MNbt} be upper and lower bounds on the usage of resource b in period t.

$$R_{MNbt} \leq r_{bt} \leq R_{MXbt}, \forall b, t \tag{9}$$

The resource variable, the resource balance and the bounds are only defined for periods where there is at least one bound on the resource variable.

4.5 Transportation and Market Constraints

To get a good enough description of the transportation capacities in operational models, one need to model pressure explicitly. For the long-term planning described here, most petroleum companies were happy with models with fixed quantity constraints in transportation pipes. If two or more pipes were built between the same nodes, these models would not account for different flows in different pipes. This means that it is the pipe paths, k, which really matters here.

When the planning period started, there was a capacity, C_{Akst}, for product s along path k in period t. Project j started in period u had a new capacity, C_{jstu}, for product s in period t. With these constants and a new variable, f_{kst}, for the flow of product s along path k in period t, the pipe path capacities were written:

$$f_{kst} - \sum_j \sum_u C_{jstu} x_{ju} \leq C_{Akst}, \forall k, s, t \tag{10}$$

The constraints are only defined from the first period when it is possible to send product s along path k. The summation over j is only for projects that expand the capacity along path k for product s. If none such j exists, the constraint is implemented as an upper bound.

Normally, there are few markets compared with the number of nodes, but even so the model is constructed in such a way that it is possible to have a market in every node. The variable for the amount of product s delivered to the market at node n in period t, is written m_{nst}. This is modelled as a delivery from the node. S_{Ink} is equal to $+1(-1)$ if node n is an end node (a start node) for path k.

$$\sum_j p_{jst} + \sum_k S_{Ink} f_{kst} - m_{nst} = 0, \forall n, s, t \tag{11}$$

These constraints (11) say that the flow into a node is equal to the flow out of the same node in every period. The summation over j is only for projects that deliver their product directly to node n. The flow variables account for the flow from and to other nodes, while the market variables account for product leaving the transportation system.

The upper and lower bounds for delivery of product s to market n in period t, M_{MXnst} and M_{MNnst}, are given as user input for individual time periods, and modelled as bounds.

$$M_{MNnst} \leq m_{nst} \leq M_{MXnst}, \forall n, s, t \tag{12}$$

4.6 The Objective

The model is written such that the user can choose between two objectives. It is possible either to minimise a weighted sum of deviations from a given goal on production or resource usage, or to maximise the total net present value from all the projects. Only the maximisation of the net present value will be discussed. From the original data such as production profiles, cost profiles, product prices and interest rates, we can calculate the contribution to the net present value from each project. This calculation is not described here.

The net present value, N_{Xju}, for project j started in period u is calculated without taking care of the contribution from any production. The net present value, N_{Pjst}, of one unit of product s produced from project j in period t is calculated in the same way. For all the markets, the user needs to specify a unit price for all products in every period. The price paths for one of the markets are called reference prices, and these are used in the calculations of N_{Pjst}. To get the correct total net present value, we also need to calculate the change in the total net present value, N_{Mnst}, for delivering a unit of product s to market n in period t, caused by the difference in prices between this market and the reference market.

This means that we can write the objective in the following way:

$$Max\ z = \sum_j \sum_u N_{Xju} x_{ju} + \sum_j \sum_s \sum_t N_{Pjst}\, p_{jst} + \sum_n \sum_s \sum_t N_{Mnst}\, m_{nst} \tag{13}$$

The first summation over j is for all j, because all projects are expected to have at least some of their costs fixed to the start of the project. The next summation over j is only for projects with production.

In the first version of the model all production was fixed and there was only one market. Then, the value of the production was put into N_{Xju} and only the first term of the objective was present. When flexible production was introduced, the objective got the second term. With several markets, the last term got included.

5 Stochastic Development Planning

During the 1980s, the price of raw oil dropped significantly, making a new kind of price-sensitivity awareness within the industry. As a consequence, some need for further analytic approaches including stochasticity, primarily on price/demand was introduced.

Professor Jörnsten, contributing significantly to the introduction of formal deterministic mathematical programming modelling within the Norwegian petroleum industry, was perhaps also among the first addressing possible price uncertainty problems related to field and transportation planning on the Norwegian continental shelf. In Jörnsten (1992), Professor Jörnsten introduces stochastic integer programming methods to Norwegian petroleum companies applying the Progressive Hedging algorithm on a scenario based formulation. Here, he shows promising computational opportunities adding the Pivot and Complement heuristic to the resulting large-scale zero-one mathematical programme.

At around the same time, the previously mentioned research group a SINTEF in Trondheim also got some projects both from Norwegian Petroleum Directorate (NPD) and STATOIL related to similar problems. This work led to the PhD-thesis (Haugen 1991) for one of the authors, with the other author as the supervisor. This contribution was perhaps more related to the computationally efficient use of modern vector processing, but it still contains some relevant stochastic modelling related to price/demand uncertainty for the Norwegian petroleum industry. A later paper by Haugen (Haugen 1996) focused more qualitatively on resource uncertainty. Both these latter approaches utilized Stochastic Dynamic Programming as the solution method.

In principle, these modelling attempts tried to add stochasticity to the type of model described in Sect. 4. These models, themselves being hard enough to solve, indicate that the attempts to include uncertainty had to contain simplifications. The number of possible projects – as well as time periods – was typically significantly reduced compared to the full-scale cases treated deterministically by companies.

We choose not to go into greater modelling details on these models. It seems however interesting to point out that all the three above discussed models were company initiated and was financed by the industry – at least initially. This is by itself interesting, not many other industries initiate stochastic programming development. The typical situation – at least in our experience – is the opposite.

It should also be pointed out that these early stochastic modelling attempts did not survive in the companies as operative models. Whether this was due to added (stochastic) informational needs, necessary simplifications compared to the deterministic models or a more or less steadily increasing price of oil and natural gas, we shall not judge. However, the price drop of oil in the 1980s was definitely a releasing factor for the company initiated projects leading to the above described academic/scientific development.

6 A Surprisingly Flexible Modelling Environment

Over a period of time, different investment planning issues have been addressed by regulating authorities (NPD) as well as companies. The agents themselves indicate that these models have been helpful through various epochs and political regimes.

It is perhaps safe to say that the scientific personnel saw the importance of building flexible models early. It is obvious that this flexibility has proven valuable in the lifetime of these models regarding their ability to 'grow' and develop. The relatively generic model structure built on the 'project concept' discussed in Sect. 4 may be seen as a key to the models' use through all these years. The model concepts' of simple building blocks have proven flexible enough to cope with different structural problems. Obviously, the models have been redeveloped, but the basic structure has in many perspectives been relatively constant. As such, the redevelopment costs have been kept relatively small, keeping these models alive.

In the following section, we will indicate this model flexibility by a short discussion of some of the more important topics where such models have played a role.

7 Practical Model Usage

It is interesting to note that the early political discussion has focused mainly on production level and to some extent on the level of annual capital investment in the petroleum sector. The (early) focus has to a limited extent been on the income generation.

The discussion on whether to plan for 40, 60 or 90 mill tons of annual petroleum productions was raised mainly due to concern with the impact on the rest of the society and due to concern with the depletion rate of an exhaustible resource.

Looking back on NPD's investment planning, NPD also tended to focus much on the use of input factors like annual spending on capital goods, the demand for construction and engineering man-hours and the use of other skilled personnel. With the new modelling capabilities NPD had ample possibilities to play with different constraints regarding input factors.

The model also gave NPD the possibility to choose an objective for the model where the weighted sum of deviations from a given goal was minimized. This way NPD could find potential sequences of fields, which gave low annual variations in

the use of specific input factors, thereby reducing the shocks on other sectors of the economy. By comparing the net present value of such solutions with solutions optimized under normal constraints for the same input factor and with the objective to maximize net present value, it was possible to calculate the cost for obtaining a more even level of demand for special input factors.

More recently, NPD has used the model to calculate the total value of the petroleum resources. In this calculation it was necessary to find a sequence for the development of future fields and prospects, and calculate the value of the cash flows.

8 What About Today?

Up to now, we have discussed the use of mathematical programming development planning models into the mid-1990s. It seems important to point out that these model concepts are still very much alive in Norwegian oil companies as well as within regulating authorities. The fact that a group from SINTEF together with StatoilHydro and GASSCO are nominated to the Franz Edelman finals this year (2008) for projects applying mathematical programming methods (see GASSOPT (Tomasgard et al. 2007; Midthun 2007)), should indicate this. These models are perhaps slightly different, focusing a bit more on operative decisions. Still, to our knowledge, models relatively equal to the models described in this chapter are still routinely used both by major Norwegian private and public agents in the sector. The models are of course more advanced both from a computational point of view as well as the technological 'look' through faster and more versatile computing platforms. Still, the initial work performed by the groups at CMI and SINTEF in the early 1980s seems to have had an important impact even today, close to 30 years later.

So, what about the future? As Fig. 1 indicates, it may seem that the need for field sequencing is diminishing due to the sheer lack of fields. However, one should bear in mind that recent price patterns indicate that previously discovered fields – judged unprofitable at the time – may very well change status. At the same time, technological development will go on, opening up new opportunities for added production from old fields. Finally, certain previously unexplored areas on the Norwegian continental shelf previously being kept off the industry, like substantial promising areas outside Lofoten, these days are being politically more 'possible' than before. This is perhaps not good news for environmentalists, on the other hand, it may open up for the next 30 years, petroleum production on the Norwegian continental shelf – keeping demand high and steady for continued use of mathematical programming tools and models.

References

Aboudi, R., Hallefjord, Å., Helgesen, C., Helming, R., Jörnsten, K., Pettersen, A. S., Raum, T., & Spence, P. (1989). A mathematical programming model for the development of petroleum fields and transport systems. *European Journal of Operational Research, 43*(1), 13–25

Aronofsky, J. S., & Williams, A. C. (1962). The use of linear programming and mathematical models in underground oil production. *Management Science, 8*(4), 394–407

Beale, E. M. L. (1983). A mathematical programming model for the long-term development of an off-shore gas field. *Discrete Applied Mathematics, 5*, 1–9

Bodington, C. E., & Baker, T. E. (1990). A history of mathematical programming in the petroleum industry. *Interfaces, 20*(4), 117–127

Haugen, K. K. (1991). *Possible computational improvements in a stochastic dynamic programming model for scheduling of off-shore petroleum fields*. Ph.D. thesis, Norwegian Institute of Technology, 7034 Trondheim, Norway

Haugen, K. K. (1996). A stochastic dynamic programming model for scheduling of offshore petroleum fields with resource uncertainty. *European Journal of Operational Research, 88*(1), 88–100

Haugland, D., Hallefjord, Å., & Asheim, H. (1988). Models for petroleum field exploitation. *European Journal of Operational Research, 37*(1), 58–72

Jörnsten, K. (1992). Sequencing offshore oil and gas fields under uncertainty. *European Journal of Operational Research, 58*(2), 191–201

McFarland, J. W., Lasdon, L., & Loose, V. (1984). Development planning and management of petroleum reservoirs using tank models and nonlinear programming. *Operations Research, 32*(2), 270–289

Midthun, K. T. (2007). *Optimization models for liberalized natural gas markets*. Ph.D. thesis, Norwegian University of Science and Technology, 7034 Trondheim, Norway

Nygreen, B., Christiansen, M., Haugen, K. K., Bjørkvoll, T., & Kristiansen, Ø. (1998). Modeling Norwegian petroleum production and transportation. *Annals of Operations Research, 82*, 251–268

Sullivan, J. (1988). The application of mathematical programming methods to oil and gas field development planning. *Mathematical Programming, 42*, 189–200

Tomasgard, A., Rømo, F., Fodstad, M., & Midthun, K. (2007). Optimization models for the natural gas value chain. In G. Hasle, K.-A. Lie & E. Quak (Eds.), *Geometric modelling, numerical simulation, and optimization* (pp. 521–558). Berlin Heidelberg: Springer

Analysis of Natural Gas Value Chains

Kjetil T. Midthun and Asgeir Tomasgard

Abstract In this paper, we provide an overview of the natural gas value chain, modeling aspects and special properties that provide challenges when doing economic analysis. We present a simple value chain optimization model and discuss important properties of this model.

1 Introduction

This paper is a tutorial on economic modeling and analysis in the natural gas value chain. The tutorial is by no means a complete overview but focuses on the most important properties (for a more detailed overview, see Rømo et al. (2006) and Midthun et al. (2008)). Special attention is given to system effects and the importance of the portfolio perspective in the value chain. This tutorial will give an understanding for the level of detail needed when doing analysis of natural gas networks.

The natural gas value chain has special properties that make it challenging to analyze. In order to describe the value chain in a mathematical model, we have to make assumptions that simplify the real world processes. There will be a tradeoff between representing details in the physical properties and creating a solvable model which can be used for analysis. In this paper, we discuss some of these tradeoffs and present our view on the properties that are essential for doing analysis. We also discuss how some assumptions can drastically alter the results of economic analysis in natural gas networks.

The case study and focus for this paper will be the value chain for natural gas in the North Sea, with an emphasis on the upstream network. We define the

This research is sponsored by the Research Council of Norway through project 176089.

K.T. Midthun (✉)
Department of Applied Economics, SINTEF Technology and Society, 7036 Trondheim
e-mail: kjetil.midthun@sintef.no

A. Tomasgard
Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, 7491 Trondheim, Norway
e-mail: asgeir.tomasgard@iot.ntnu.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_5,
© Springer-Verlag Berlin Heidelberg 2010

upstream network as the production facilities and transportation infrastructure in the North Sea. The discussion also includes the market nodes in continental Europe and the UK. The downstream network in these countries are, however, not part of our discussion.

In Sect. 2, we provide a presentation of the most important model classes within the natural gas industry as well as some references. The natural gas value chain is introduced in Sect. 3, before we present a simple mathematical model for value chain optimization in a natural gas network in Sect. 4. The model is then extended and important characteristics of the natural gas value chain are discussed. In Sect. 5, we give some conclusions and suggestions for interesting topics for future research.

2 Literature Review

The petroleum industry has been a pioneer in the application of operations research, and the literature is therefore extensive. In Bodington and Baker (1990), an interesting overview of the history of mathematical programming in the petroleum industry is given. We present a short overview of some of the most important model classes; investment models, value chain models, transportation models and equilibrium models.

Investment models give decision support for strategic decisions such as field investments and sequencing of investments. There are a large number of publications within this field. This is not surprising given the large risks and costs associated with offshore investments. There exist a number of deterministic investment models, such as Sullivan (1988), Haugland et al. (1988), Nygreen et al. (1998) and van den Heever et al. (2001). There are also some models which incorporate uncertainty, such as Jörnsten (1992), Haugen (1996), Jonsbråten (1998), and Goel and Grossmann (2004). The uncertain parameters in the models include future demand for natural gas, development of oil prices and available reserves in the fields.

The special properties of the transportation network make a value chain approach to optimizing the system important. In the value chain approach, the complete network is considered and optimized simultaneously (the special properties of the transportation network are discussed later in this paper). The value chain approach has become even more valuable after the liberalization process, which meant an increase in flexibility for the participants in the value chain. Examples of value chain models are Ulstein et al. (2007), Selot et al. (2007), Tomasgard et al. (2007) and Midthun et al. (2007a).

The transportation of natural gas is one of the key elements when studying the natural gas industry. Because of the interdependence among flows in pipelines, it is important to find a tradeoff between accurately describing the properties of the transportation network, and being able to solve the model. A simplified representation leads to an inaccurate model of the transportation (and may lead to wrong conclusions), while a too detailed presentation makes the model non-linear and non-

convex. Examples of transportation models with emphasis on the physical properties are De Wolf and Smeers (2000), O'Neill et al. (1979), Westphalen (2004) and Selot et al. (2007).

Equilibrium models are used to study situations where more than one player acts strategically. The models are formulated as complementarity problems. A good overview of complementarity problems in natural gas markets is given in Gabriel and Smeers (2005). The paper gives a survey of some of the existing models, as well as develops relevant models for the restructured natural gas markets. Other examples of equilibrium models with application in natural gas are De Wolf and Smeers (1997), Boots et al. (2004), Gabriel et al. (2005) and Zhuang and Gabriel (2006).

3 The Natural Gas Value Chain

Natural gas is formed naturally from plant and animal remains. Subjected to high pressure and temperature over millions of years, the organic material changed into coal, oil and natural gas. The natural gas can be found in porous geological formations (reservoirs) beneath the earth's surface. In these reservoirs, the gas can be in gaseous phase or in solution with crude oil.

A simplified picture of the offshore natural gas value chain in the North Sea is shown in Fig. 1. The gas is transported from the production fields to processing plants, or directly to the market hubs in Europe. There are storage possibilities along the transportation route. In addition, the transportation network itself can be considered as a storage facility since there are large volumes of gas contained in the pipelines at all times. We give below a short presentation of the most important elements in the value chain.

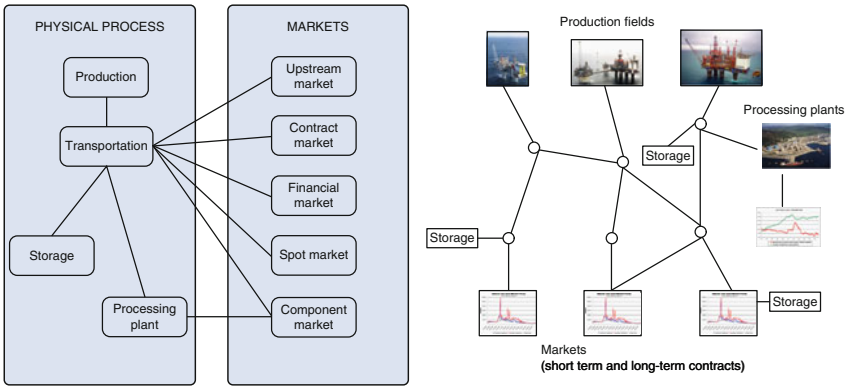


Fig. 1 Illustration of the natural gas value chain

3.1 Production and Processing

The gas is produced from the reservoirs. The driving forces are pressure from the expanding gas as well as water which causes the gas to flow into the well. The gas production depends on the pressure in the reservoirs. High pressure in the reservoir gives a high production rate. In order to increase the pressure in the reservoir, and thus increase production capacity, compressors are sometimes used.

The natural gas sold to Europe consists mainly of methane (dry gas). The gas produced at the fields can, however, contain other components with market value, such as associated hydrocarbons. Gas containing both dry gas and associated hydrocarbons is called rich gas. The rich gas is transported to processing plants where the dry gas and wet gas (the associated hydrocarbons) are separated. The wet gas is then heated in order to separate the different components which, in turn, are sold in component markets.

3.2 Transportation

In the North Sea, the gas is transported in long, subsea pipelines operated at high pressure levels. The gas molecules flow in the pipeline from high pressure points to low pressure points. At the production fields, pressure is increased with compressors in order to create a pressure difference that is sufficient for the gas to flow to the landing points. With the completion of Langeled, the network will consist of 7,800 km of pipelines. For details on the infrastructure and topology in the North Sea, see OED (2007).

3.3 Storage

The demand for natural gas shows strong seasonal patterns and large short-term volatility. Both these factors give a large value to optimal storage utilization. There are many different forms of storages that are used for storing natural gas: abandoned oil and gas fields, aquifers, LNG-storages and salt caverns. The storages are different with respect to capacity, injection and extraction capabilities and cost of operation. For more information on storages, see EIA (2002). In addition, the pipeline network can also be used as storage (line pack).

3.4 Markets

Traditionally, the gas from the North Sea has been sold in long-term take-or-pay contracts (TOP). In the TOP-contracts, the price is determined based on a

formula whose main components are the prices of competing fuels (for instance oil). A yearly volume is decided, and then the buyers have flexibility with respect to nomination on shorter time periods (within certain limits). One of the results from the ongoing liberalization process in the European gas industry is emerging short-term markets for natural gas.

3.5 Mathematical Models

In this section, we present a simple model for a transportation network for natural gas. The natural gas value chain can be modeled as a collection of nodes: fields, junction points and markets. In addition, we need pipelines to connect the nodes in a transportation network. The gas molecules flow in the pipelines from high pressure nodes to low pressure nodes. The volume of gas that flows in a pipeline between an inlet and an outlet point is dependent on the pressure difference between these two points and the design parameters of the pipeline. The design parameters incorporate, amongst others, the length and diameter of the pipeline. To relate the design parameters and the pressure difference to the actual flow in the pipelines, the Weymouth equation (see, for instance, Katz and Lee (1990) can be used.

There are various choices for an objective function for the model, such as maximize flow, minimize costs, maximize profits and maximize social surplus. In our model, we maximize a utility function $U(k_g, r_i, f_{ij})$. The variables in the model are the production k_g in each field g, the pressure r_i in node I, and the flow f_{ij} between nodes i and j. The objective function can then be formulated as:

$$\max_{k_g, r_i, f_{ij}} U \tag{1}$$

We then add constraints to take care of production limits, capacity limitations in the pipelines, pressure limits in the nodes, mass conservation in the network and demand satisfaction in the market nodes. The first set of constraints ensures that mass is conserved in the network. Production k_g in node g must equal the amount of gas f_{gj} transported from the production node g into nodes j in its set of downstream nodes $O(g)$:

$$\sum_{j \in \mathcal{O}(g)} f_{gj} = k_g, \; g \in \mathcal{G} \tag{2}$$

where G is the set of all field nodes. For the junction nodes, the amount of gas that flows into node j must also flow out of node j:

$$\sum_{i \in \mathcal{I}(j)} f_{ij} = \sum_{n \in \mathcal{O}(j)} f_{jn}, \; j \in \mathcal{T}, \tag{3}$$

where T is the set of all junction nodes. In the market node m, we need to make sure that the quantity of gas delivered does not exceed the demand in the node, D_m denotes the demand:

$$\sum_{j \in \mathcal{I}(m)} f_{jm} \leq D_m, \; m \in \mathcal{M}, \tag{4}$$

where M is the set of all market nodes. Moreover, we need to make sure that the maximum and minimum requirements for the pressure in the nodes are satisfied:

$$r_i \geq \underline{R_i}, \quad i \in \mathcal{N}, \tag{5}$$

$$r_i \leq \overline{R_i}, \quad i \in \mathcal{N}, \tag{6}$$

where N is the set of all nodes in the network. The capacity in the pipelines is determined based on the Weymouth equation:

$$f_{ij} = K_{ij}^W \sqrt{r_i^2 - r_j^2}, \tag{7}$$

where f_{ij} is the flow between the inlet point in node i and the outlet point in node j, K^W is a constant determined by the design parameters of the pipeline and r is the pressure in the respective node. As we can see from this expression, the relation between flow and pressure is not linear. Actually, it describes a quarter cone (see Fig. 2).

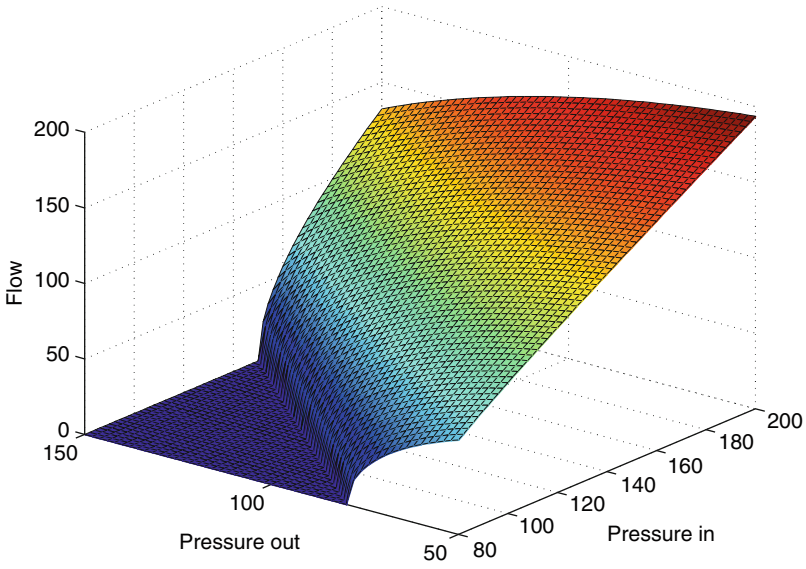


Fig. 2 The Weymouth equation relates the design parameters in the pipeline and the pressure in the entry and exit nodes to the flow in the pipeline

In order to keep the transportation model as an LP model (and thus enable analysis in large-scale networks), we use a linearized version of the Weymouth equation. The linearization is based on a first-order Taylor series expansion around a set of fixed pressure levels (RI,RO):

$$f_{ij} \leq K_{ij}^{W} \frac{RI_{il}}{\sqrt{RI_{il}^2 - RO_{jl}^2}} r_i - K_{ij}^{W} \frac{RO_{jl}}{\sqrt{RI_{il}^2 - RO_{jl}^2}} r_j, i \in \mathcal{N}, j \in \mathcal{O}(i), l = 1, \ldots, L, \tag{8}$$

where we use L linear constraints for each pipeline. For more details on the linearization, see Rømo et al. (2006).

3.6 System Effects

In natural gas transportation networks with several nodes and with more than one pipeline connected to at least one of the nodes, there are system effects in the network. The system effects come from the influence the pressure in a node has on the flow in connected pipelines. If two pipelines are connected to the same node, a pressure increase in the node will influence the potential flow in both pipelines. Depending on the topology of the network, the dependencies between pipelines may be strong. Consider the small network in Fig. 3. The pressure in junction node j_1 influences the flow in all the three pipelines in the network. An increase of the pressure in node j_1 will decrease the capacity from field g_1 to j_1, and increase the capacity from j_1 to both market m_1 and m_2. Even in this very small network, the system effects may be important, and it is difficult to determine appropriate fixed capacities for the pipelines.

Midthun et al. (2008) discuss the importance of system effects for both simple maximize flow formulations as well as more complicated economical objective functions. In electricity, various papers concerning externalities in the network exist

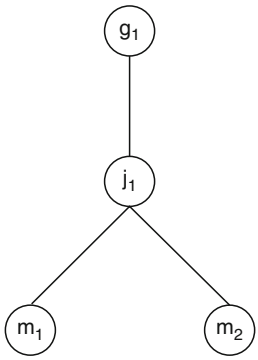


Fig. 3 A simple transportation network consisting of a field node, a junction node and two market nodes

(Bjørndal 2000; Wu et al. 1996; Schweppe et al. 1988). In the natural gas literature, there are many examples of papers that present technical models of the transportation problem (Ehrhardt and Steinbach 2005; Martin et al. 2006; De Wolf and Smeers 2000; O'Neill et al. 1979; Westphalen 2004), as well as economical models with a simplified representation of the transportation networks (Cremer and Laffont 2002; Cremer et al. 2003).

3.7 Markets and the Portfolio Perspective

The development of short-term markets adds flexibility for the producers of natural gas. In addition to creating new market possibilities (delivery of additional volume of gas) and risk management tools (financial markets), the short-term markets also enables the producers to do virtual routing of gas (in space and time). For our mathematical model, the inclusion of short-term markets enables us to change (4) to the following, more flexible, version:

$$\sum_{j \in \mathcal{I}(m)} f_{jm} = D_m + q_m, m \in \mathcal{M}, \tag{9}$$

where q_m is the traded volume in the short-term market. This traded volume can be both positive and negative. This means that the producers now have the possibility to supply the demand in the market by purchasing gas in a spot market, as well as the possibility to sell additional gas in the spot market.

To illustrate how this flexibility makes a portfolio perspective vital, consider the network in Fig. 4. In the situation before the liberalization, a TOP-contract would specify, for instance, that field g_1 should deliver to market m_1 while g_2 should deliver to market m_2. After the liberalization, the companies in the respective fields and contracts can freely choose which field should deliver to which market node. When short-term markets are introduced to this system, the flexibility in the system is drastically increased. In a situation where field g_1 needs to produce, and there is a demand in a TOP-contract in market m_1, the delivery can be made in several

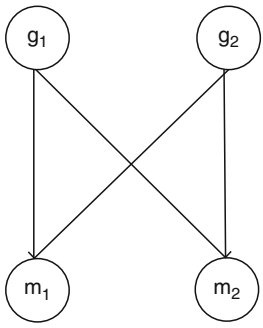


Fig. 4 An example network

ways. First of all, one possibility is to let field g_1 produce and then ship the volume to market m_1. Another alternative is to send the volume produced in field g_1 to the spot market in market m_2, and then buy spot in m_1 to deliver in the TOP-contract. Alternatively, field g_2 can produce and deliver in market m_1 while the production from g_1 is sent to market m_2. If storages are introduced in the system, the possibilities naturally increase even further. Also, in a large natural gas network there will be a very large number of possible combinations that let field g_1 produce and let the company fulfill the agreements in the TOP-contract in market m_1.

The discussion of routing in time is analogous. Consider a situation where field g_1 needs to produce in time T, while the company must deliver in a TOP-contract in market m_1 in time $T+1$. One way of doing this is to send the production from g_1 to storage in time T, and then send the gas from storage to market m_1 in time $T+1$. Alternatively, the production in g_1 can be sold in a spot market in time T, and in time $T+1$; either g_2 can produce and send the gas to market m_1, or the company can buy spot in m_1 to deliver in the TOP-contract. A lot of different possibilities are open to the producer also in this case.

3.8 Modeling Competition

Normally, each production field has several license owners. Owing to the European Union regulations as specified in the Gas Directive (European Union 1998), these companies have to sell their gas independently. The transport capacity needed to get the gas to the market is offered as an independent service and is administered by the company Gassco. For a detailed description of this system and a more precise model of the competition that arise in the transport markets, see Midthun et al. (2007b). Here, we will give a simplified example of how our previous model can be extended to handle competition. Naturally, the complexity increases as we move from the model class of linear programming to modeling Nash equilibrium as a complementarity problem. Some other references to similar competition models are Hobbs (2001) and Wei and Smeers (1999).

We need to define some new notation based on the model in Sect. 4. Now, f_{lgj} is the flow from producer l in production node g into nodes j in its set of downstream nodes $O(g)$. Similarly, f_{ljm} is the flow from producer l into market node m from any of its upstream nodes $I(m)$.

The mathematical model for producer l is then:

$$\max_f \sum_{m\in\mathcal{M}} \sum_{j\in\mathcal{I}(m)} p_m f_{ljm} - \sum_{g\in\mathcal{G}} \sum_{j\in\mathcal{O}(g)} \lambda_{gj} f_{lgj} - \sum_{m\in\mathcal{M}} \sum_{j\in\mathcal{I}(m)} \lambda_{jm} f_{ljm} - \sum_{g\in\mathcal{G}} C_g(k_{lg}). \tag{10}$$

Here, $C_g(k_{lg})$ is a convex production cost function in node g for player l. The production level is given by k_{lg}. Also, λ_{gj} is the entry tariff for each unit inserted in the network in pipeline g_j and λ_{jm} is the exit tariff for each unit withdrawn from

the network in pipeline j_m. Here, p_m is the price of natural gas in a competitive spot market m. In addition, we must make sure that the production in each field is the same as the producer's flow out of the field node and that each producer's sales are in accordance with his production:

$$\sum_{j \in \mathcal{O}(g)} f_{lgj} = k_{lg}, \quad l \in \mathcal{L}, g \in \mathcal{G} \tag{11}$$

$$\sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{I}(m)} f_{ljm} = \sum_{g \in \mathcal{G}} k_{lg}, l \in \mathcal{L}. \tag{12}$$

We must also model the optimization problem for the independent system operator (ISO) managing the network service. A number of objective functions for the network operator may be considered, see, for example, Midthun et al. (2007b) for a discussion. Here, we choose to maximize the tariff income in an entry–exit system. Note that the decision variables of the system operator are the routing of natural gas and allocation of capacity to the producers, while the tariff λ (see (14) and (15)) is a consequence of these decisions. The motivation is as follows: by choosing the dual variables as the tariff (it is not a decision variable for the system operator), each of the producers have 0 profit on their last produced unit while the routing of the system operator is used to maximize his tariff income:

$$\max_f \sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{I}(m)} \lambda_{jm} f_{jm} + \sum_{g \in \mathcal{G}} \sum_{j \in \mathcal{O}(g)} \lambda_{gj} f_{gj}. \tag{13}$$

The optimization problem for the ISO consists of this objective function and (3), (5), (6) and (8) from the original formulation.

In addition, we add market-clearing conditions to the complementarity problem. These conditions determine the tariff in the system, and are given by the following equations (L is the set of all producers l):

$$\sum_{j \in \mathcal{O}(g)} f_{gj} = \sum_{j \in \mathcal{O}(g)} \sum_{l \in \mathcal{L}} f_{lgj}, g \in \mathcal{G}, \tag{14}$$

summarizing the entry flow into the network in each production node and

$$\sum_{j \in \mathcal{I}(m)} f_{jm} = \sum_{j \in \mathcal{I}(m)} \sum_{l \in \mathcal{L}} f_{ljm}, m \in \mathcal{M}, \tag{15}$$

summarizing the exit flow into each market node. The dual variables for these constraints give, respectively, the entry tariff λ_{gj} and the exit tariff λ_{jm}.

Under the assumption that all the decisions are made simultaneously, this is a complementarity program. It is normally solved by stating the Karush–Kuhn–Tucker conditions (Karush 1939; Kuhn and Tucker 1951) of all the players and solving the total system of equations, for example, using a solver like PATH (Dirkse and Ferris 1995).

4 Conclusions

We have presented a short tutorial on modeling of the natural gas value chain. The various elements in the value chain have been introduced, and a simplified mathematical model that can be used for analysis has been presented. We have stressed the importance of keeping a portfolio perspective when planning and analyzing the value chain. In addition, we advocate that modeling of the flow-pressure dependencies in natural gas transportation networks is vital. If fixed capacities are assumed on the pipelines, the dynamics of the system is lost.

References

Bjørndal, M. (2000). Topics on electricity transmission pricing (PhD thesis, Norwegian School of Economics and Business Administration, Bergen).

Bodington, C. E., & Baker, T. E. (1990). A history of mathematical programming in the petroleum industry. *Interfaces, 20*(4), 117–127.

Boots, M. G., Rijkers, F. A. M., & Hobbs, B. F. (2004). Trading in the downstream European gas market: a successive oligopoly approach. *The Energy Journal, 25*(3), 73–102.

Cremer, H., Gasmi, F., & Laffont, J. (2003). Access to pipelines in competitive gas markets. *Journal of Regulatory Economics, 24*(1), 5–33.

Cremer, H., & Laffont, J. (2002). Competition in gas markets. *European Economic Review, 46* (4–5), 928–935.

De Wolf, D., & Smeers, Y. (1997). A stochastic version of a Stackelberg-Nash-Cournot equilibrium model. *Management Science, 43*(2), 190–197.

De Wolf, D., & Smeers, Y. (2000). The gas transmission problem solved by an extension of the simplex algorithm. *Management Science, 46*(11), 1454–1465.

Dirkse, S. P., & Ferris, M. C. (1995). The PATH solver: a non-monotone stabilization scheme for mixed complementarity problems. *Optimization Methods and Software, 5*, 123–156.

Ehrhardt, K., & Steinbach, M. (2005). Nonlinear optimization in gas networks. In H. E. Bock, (Ed.), *Modeling, simulation and optimization of complex processes* (pp. 139–148). Berlin, Heidelberg, New York: Springer.

EIA. (2002). The basics of underground natural gas storage. Resource document. Energy Information Administration. http://www.eia.doe.gov/pub/oil_gas/natural_gas/analysis_ publications/storagebasics/storagebasics.html. Accessed 23 March 2009.

European Union. (1998). Directive 98/30/EC of the European parliament and of the council.

Gabriel, S., Kiet, S., & Zhuang, J. (2005). A mixed complementarity-based equilibrium model of natural gas markets. *Operations Research, 53*(5), 799–818.

Gabriel, S., & Smeers, Y. (2005). *Complementarity problems in restructured natural gas markets*. CORE Discussion Paper No. 2005/37, Center for Operations Research and Econometrics, Catholic University of Louvain, Belgium.

Goel, V., & Grossmann, I. E. (2004). A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers and Chemical Engineering, 28*, 1409–1429.

Haugen, K. K. (1996). A stochastic dynamic programming model for scheduling of offshore petroleum fields with resource uncertainty. *European Journal of Operational Research, 88*, 88–100.

Haugland, D., Hallefjord, Å., & Asheim, H. (1988). Models for petroleum field exploitation. *European Journal of Operational Research, 37*(1), 58–72.

Hobbs, B. E. (2001). Linear complementarity models of Nash-Cournot competition in bilateral and POOLCO power markets. *IEEE Transaction on Power Systems, 16*(2), 194–202.

Jonsbråten, T. W. (1998). Oil field optimization under price uncertainty. *The Journal of the Operational Research Society, 49*(8), 811–818.

Jörnsten, K. O. (1992). Sequencing offshore oil and gas fields under uncertainty. *European Journal of Operational Research, 58*(2), 191–201.

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. (MSc Dissertation. Department of Mathematics, University of Chicago, Chicago, Illinois).

Katz, D., & Lee, R. (1990). *Natural gas engineering*. New York: McGraw-Hill.

Kuhn, H. W., & Tucker, A. W. (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium* (pp. 481–492). Berkeley: University of California Press.

Martin, A., Möller, M., & Moritz, S. (2006). Mixed integer models for the stationary case of gas network optimization. *Mathematical Programming, 105*(2–3), 563–582.

Midthun, K. T., Bjørndal, M., Smeers, Y., & Tomasgard, A. (2007a). Capacity booking in a transportation network with stochastic demand and a secondary market for transportation capacity. In K. T. Midthun (Ed.), *Optimization models for liberalized natural gas markets* (pp. 147–185). (PhD thesis, NTNU, 2007:205. Trondheim, Norway).

Midthun, K. T., Bjørndal, M., & Tomasgard, A. (2008). *Modeling optimal economic dispatch and system effects in natural gas networks*. Working paper, NTNU, Trondheim, Norway. Submitted to international journal.

Midthun, K. T., Nowak, M. P., & Tomasgard, A. (2007b). An operational portfolio optimization model for a natural gas producer. In K. T. Midthun (Ed.), *Optimization models for liberalized natural gas markets* (pp.113–143). (PhD thesis, NTNU, 2007:205. Trondheim, Norway).

Nygreen, B., Christiansen, M., Bjørkvoll, T., Haugen, K., & Kristiansen, Ø. (1998). Modeling Norwegian petroleum production and transportation. *Annals of Operations Research, 82*, 251–268.

OED. (2007). Fakta - Norsk petroleumsverksemd 2007 (In Norwegian).

O'Neill, P., Williard, M., Wilkins, B., & Pike, R. (1979). A mathematical programming model for allocation of natural gas. *Operations Research, 27*(5), 857–873.

Rømo, F., Fodstad, M., Tomasgard, A., Hellemo, L., & Nowak, M. (2006). *An optimization system using steady state model for optimal gas transportation on the Norwegian continental shelf*. Technical report, SINTEF, Trondheim, Norway.

Schweppe, F., Caramanis, M., Tabors, R., & Bohn, R. (1988). *Spot pricing of electricity*. Norwell, Massachusetts: Kluwer.

Selot, A., Kuok, L. K., Robinson, M., Mason, T. L., & Barton, P. I. (2007). A short-term operational planning model for natural gas production systems. *AIChE Journal, 54*(2), 495–515.

Sullivan, J. (1988). The application of mathematical programming methods to oil and gas field development planning. *Mathematical Programming, 42*(1–3), 189–200.

Tomasgard, A., Rømo, F., Fodstad, M., & Midthun, K. T. (2007). Optimization models for the natural gas value chain. In G. Hasle, K. A. Lie, & E. Quak (Eds.), *Geometric modelling, numerical simulation and optimization: applied mathematics at SINTEF*. Springer.

Ulstein, N., Nygreen, B., & Sagli, J. (2007). Tactical planning of offshore petroleum production. *European Journal of Operational Research, 176*(1), 550–564.

van den Heever, S. A., Grossmann, I. E., Vasantharajan, S., & Edwards, K. (2001). A Lagrangean decomposition heuristic for the design and planning of offshore hydrocarbon field infrastructures with complex economic objectives. *Industrial and Engineering Chemistry Research, 40*(13), 2857–2875.

Wei, J. -Y., & Smeers, Y. (1999). Spatial oligopolistic electricity models with Cournot generators and regulated transmission prices. *Operations Research*, *47*(1), 102–112.

Westphalen, M. (2004). *Anwendungen der stochastischen optimierung im stromhandel und gastransport*. (PhD thesis, University Duisburg-Essen, Germany).

Wu, F., Varaiya, P., Spiller, P., & Oren, S. (1996). Folk theorems on transmission access: proofs and counterexamples. *Journal of Regulatory Economics, 10*(1), 5–23.

Zhuang, J., & Gabriel, S. (2006). *A complementarity model for solving stochastic natural gas market equilibria*. Accepted to *Energy Economics* in January 2006.

On Modeling the European Market for Natural Gas

Lars Mathiesen

Abstract Several features may separately or in combination influence conduct and performance of an industry, e.g., the numbers of sellers or buyers, the degree of economies of scale in production and distribution, the temporal and spatial dimensions, uncertainty about long run development of demand in particular combined with large investments in production capacity and infrastructure, etc. Our focus is modeling in order to obtain insight into market mechanisms and price formation. In particular, we demonstrate the rather different solutions obtained from the price-taking behavior versus the oligopolistic Cournot behavior when the spatial dimension is observed.

1 Introduction

The conduct and performance of an industry are characterized by several features in combination, e.g., the nature of its products, the numbers of sellers or buyers, the degree of economies of scale in production and distribution, the temporal and spatial dimensions, etc. In any specific analysis, it is advisable to tailor-make a model to the issues in question.[1] We consider the development of a model to help understand price formation in the European natural gas market over a time horizon of the next 5–10 years. With this in mind, we will briefly review industry characteristics and related modeling issues.

First and foremost is the nature of the product. Natural gas can be considered a *homogeneous* good. Except for the low-calorific gas from Groeningen gases from other fields mix in the pipelines. Consumers have no preference for the gas from a particular supplier, which implies there will be one price. Of course, there are more dimensions to supply than the product as such. Security and flexibility of deliveries are concerns that may warrant price premiums and diversifications.

[1] Smeers (1997) reviews models and issues in the European gas market.

L. Mathiesen
Department of Economics, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: lars.mathiesen@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_6,
© Springer-Verlag Berlin Heidelberg 2010

Several features of the natural gas industry signal that *time* is important. In economic theory, resource extractive industries are typically studied from the perspective of optimal intertemporal depletion paths.[2] There are very long lead times between a decision to develop a field or build a major pipeline and deliveries to the market. Investments are enormous, whereby capital costs vastly dominate operating costs. Thus, uncertainties regarding future market conditions and the economic feasibility of the projects are considerable for all parties involved.[3] Another aspect of time is the *seasonal* pattern within a year. Some consumers have much higher demand in winter than in summer. With costly production capacity, it may be profitable to apply an average production rate and use storage to balance seasonal demand.

Space is important for at least two reasons. The distance from a major field in Russia, Algeria, or Norway to the market is large and combined with the above-mentioned uncertainties it may imply low and highly uncertain net-back prices. Also, distances between various consuming regions in Europe are considerable, whereby net-back prices to a given supplier differ considerably between regions and will influence his decision to supply these markets.

In the European gas market, there are a few large suppliers, some very influential transmission and distribution companies, as well as a few large consumers, although most consumers are small. Fewness and size imply potential *market power*, that is, ability to influence price. The extent of power and the effects from exertion of power are interesting issues for analysis. In production, fewness is related to the availability of resources; only a few countries are endowed with natural gas. Furthermore, extraction rights are allocated to only a few companies.[4] In transportation, fewness is a result of the quite substantial economies of scale with respect to pipeline dimension both in terms of laying the pipeline and its carrying capacity. Transmission and distribution networks are examples of natural monopolies. Old pipelines (or monopolistic behavior) may constrain flows and hinder some seller's access to a market and thus impede competition. Hence, modeling the actual *network* of pipelines with their tariffs and capacities may be relevant.[5]

One question is who has power; another is how to model market power. Consider the stylized structure in Fig. 1, where one or more gas producers sell to a transmission or a distribution company, who resells to the final consumers. The numbered

[2] Mathiesen et al. (1987) argue that shadow prices of the vast gas resources of the larger suppliers are close to zero whereby a static approach is appropriate from a resource perspective. The large profits in gas production are oligopolistic rents that stem from uneven distribution of the resources and not scarcity.

[3] In order to reduce the risks caused by long lead times and large uncertainties, parties involved write long term contracts for deliveries. The terms of such contracts may not match price formation in an equilibrium model. But it seems that contracting clauses rather than being written in stone are modified when market conditions change, thereby adapting to the logic of the market mechanism.

[4] Algeria, Russia, and The Netherlands each have only one (large) producer, while in Norway a central board coordinated sales for years.

[5] For the analysis of investment in pipe capacities, modeling of flow as dependent on pressure at nodes may be important (see De Wolf and Smeers 1993).

Fig. 1 Combinations of market power

arrows represent exertion of market power. 1, for example, signals that producers exploit market power versus price-taking transmitters and 3 represents the opposite, when a transmission company exploits its power versus price-taking producers. While one may observe that several of these relationships coexist in an industry, economic theory makes the modeling of some combinations difficult. The problem is whether the theory provides a (locally) unique solution or not.

Within *noncooperative* game theory, it is well known how to model 1, namely, an oligopoly of producers selling to price-taking transmitters. This is the Nash–Cournot model. 3 is that model turned upside down, i.e., an oligopsony of transmission companies buying from price-taking producers. Although one does not see many applications, it is conceptually sound. Combination 1&2 describes a situation of successive market power, where producers sell to price-taking transmitters, who in turn exploit their power over price-taking customers. Combination 2&3 is about a company that has market power in both factor and product markets. It can be generalized to an oligopolistic setting.

Combination 1&3 signals a situation where each side of a market has power. Conceptually, this case is similar to a so-called bilateral monopoly. Rather than having a unique solution, it involves a continuum of solutions. The impasse may be resolved by using the Nash bargaining solution concept of *cooperative* game theory. This theory, however, is not as easily available as the noncooperative theory for large, detailed, numerical models. Thus, one has to decide which features are essential to model, and which are not. Our model concept is based upon noncooperative game theory, whereby we cannot model the explicit exertion of market power of two opposing agents in a market. There are indirect ways to analyze the realities of 1&3 though, and we return to such analysis below.

The above mentioned features, in addition to the number of details in each dimension, are relevant for some analyses. Even though present day PCs have an enormous computing capacity, one has in practice, however, to abstain from bringing in every dimension in minute detail. If not for other reasons, one should acknowledge that such a model is meant to support rather than make decisions, and therefore has to be transparent enough for the user to understand what is going on inside the model.

The remainder of the chapter is structured as follows. Section 2 reviews some attempts at modeling natural gas markets, the focus is on behavior. Section 3 discusses the modeling of behavior in a general setting. This is embedded in a spatial model in Sect. 4, and striking differences of solutions of two different modes of behavior are illustrated. Section 5 compares two types of transportation models. Section 6 concludes the chapter.

2 Previous Modeling

An early effort is the Gas Trade Model developed in the Systems Optimization Laboratory at Stanford around 1980 (Beltramo et al. 1986). It is a multiregional transportation type model of the North American market and assumes price-taking behavior in supply and demand. Berg (1990) employed a version of the model for an analysis of increased Canadian supply to the US market, while Boucher and Smeers (1984) applied this structure for an analysis of the European market.

Observing clear indications of exertion of market power on the selling side in Europe, Mathiesen et al. (1987) studied the consequences for model solution of alternative kinds of seller behavior, namely price-taking, Nash–Cournot or collusion. They focused on the three large exporters to continental Europe – Algeria, Norway, and the Soviet Union – and two large internal suppliers – UK and The Netherlands. In addition, there was indigenous production in several of the 12 regional consumption areas. They concluded that a modified Cournot-model seemed to track observed volumes and prices best.

Using essentially the same model structure, Golombek et al. (1995, 1998) analyzed various aspects of the deregulation process of the European energy market. For example, the one Norwegian Cournot-player was replaced by several players whereby supplied volumes increased and prices dropped. Boots et al. (2004) considered the case of a successive oligopoly (the combination of 1&3 in Fig. 1) in a disaggregated market. Eldegard et al. (2001) developed a similar model and included in addition the storage of gas between seasons. All these models consider several supply and consumption regions and they are of transportation type (see Sect. 5).

Brekke et al. (1987) argued that applying a static model to analyze 10–15 years into the future seriously misrepresented the investment process over this time span. They built an intertemporal model where suppliers engaged in a dynamic game observing previous actions and reacting to them. Because of the time-dimension and added complexity in the strategy-space, they condensed other aspects. For example, they considered only one aggregate European excess demand – net of indigenous production. They disregarded the UK market and subsumed Dutch production in indigenous production, thus considering a game between the three players: Algeria, Norway, and the Soviet Union.

Bjerkholt et al. (1989) studied market power exerted by the transmission companies and possible effects of a deregulation in terms of third party access to pipelines of the European gas market by 1992. They did not formulate a model for optimal tariffs (cf., our discussion related to combination 1&3 in Fig. 1). Hoel et al. (1987) modeled a cooperative game between sellers and buyers represented by transmission companies. Grais and Zheng (1996) focused on the East–West gas trade and studied a game with Russia as a Stackelberg leader versus a transmitter (in Eastern Europe) and a Western European importer. von Hirschhausen et al. (2005) also studied Russian options of transporting gas to Western Europe using both noncooperative and cooperative strategies. Like the Brekke et al. model, these four analyses led to fairly compact models.

Mathiesen (1987) extended the Mathiesen et al. (1987) model to allow for an explicit network of transmission lines. The model was constructed for Statoil to simulate various scenarios in the European gas market. To facilitate such analyses, the model was embedded in an interactive program and presented as a decision support system. The user could choose the time horizon and modify some drivers of natural gas prices, like the prices of alternative energies as light and heavy fuel oils, growth in GDP, and the change of some energy policy parameters within a region.[6] The model observed five different types of consumers in each region, with price and income elasticities depending on the time horizon of the analysis.

Using the network model, Mathiesen (1988) demonstrated the importance of the spatial dimension in combination with strategic behavior. Particularities of the geography are lost when assuming that sale takes place at a central destination in Europe. Selling strategies for a producer may look very different in a spatial and a nonspatial model and the sales pattern of the competitive and the noncompetitive equilibria differ tremendously, as will be illustrated below. Simply because of different location and distances to the various regions of the market, Algeria may benefit from a more aggressive sales strategy than Norway and Russia, and there may be little the two can do to counteract such sales.

Tandberg analyzed the potential for gas in Scandinavia (Tandberg 1990, 1992). He studied investments in a pipeline system to supply Sweden and employed a model that considered three forms of energy. Natural gas and electricity were modeled to flow from gas fields or power plants to consumers through separate networks, while fuel oil was supplied at a unit cost. Energy demand originated in six consumer groups in twelve regions, where some consumers could entertain fuel switching dependent on relative prices. Tandberg observed that gas supplies would increase total energy supplies and reduce electricity prices – benefitting consumers – and reduce the value of extant electricity generating capacity in Sweden – hurting producers. He used the output from the equilibrium model to establish a pay-off matrix for a game between electricity producers (wholesale) and retailers (representing consumers) in Sweden.

Øygard and Tryggestad (2001) consider the deregulation of the European gas market using the McKinsey model. Judging from their nontechnical description, and a presentation by Keller (2001) presumably of the same model, it seems to be a fairly detailed network-type model with large numbers of production fields, pipelines, consuming regions, and consumer types, but where suppliers and other agents are price takers. It may seem that the behavioral dimension is sacrificed for details. This may be reasonable strategy for some kind of analyses. As seen from our numerical illustration below, however, the trade pattern of such a model, or changes in trade pattern caused by shifts in parameters, may be next to noninformative. One should appreciate the insights gained from a combination of behavior and geography (Mathiesen 1988), or behavior and time (Brekke et al. 1987), both of which combinations are not available in the McKinsey model.

[6] The Statoil model is expanded, its database is updated, and the interactive program improved. It is helpful to analysts (see Fuglseth and Grønhaug 2003), but it may also be too complicated for most non-academics.

3 Behavior

In many markets, producers have the ability to influence price and their exertion of such power is an essential feature of the market. Adopting an assumption of price-taking behavior may seriously distort model solutions. Of course, whether a particular market, like the European natural gas market, is best modeled as noncompetitive rather than a competitive one is an empirical question. Rather than delve on this question, we will illustrate some major differences in the solutions for these alternative model assumptions.

Consider the market for a homogeneous product and assume that producers make volume decisions. For ease of exposition in this section, assume there is one nonsegmented market. Let i denote producer, $i = 1, \ldots, n$,

x_i and $C_i(x_i)$ denote his volume and total cost of producing it,

$p(X)$ denote the market price as a function of total supply, i.e., $X = \Sigma_i x_i$, and

$\pi_i(\mathbf{x}) = p(X)x_i - C_i(x_i)$ denote his profits, where $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector of decision variables for all n producers.

The profit maximization problem of producer i, is then

$$maximize\, \pi_i\,(\mathbf{x}) = p\,(X)\,x_i - C_i\,(x_i)\,,\; i = 1, \ldots, n. \tag{1}$$

Through the vector $\mathbf{x}$, his maximization problem is a function of all rivals' decision variables as well. Let $c_i' \equiv \partial C_i/\partial x_i$. Provided $x_i > 0$, his first-order (necessary) condition for a profit maximum are

$$\begin{aligned}\partial\pi_i/\partial x_i =& p + [(\partial p/\partial x_1)(\partial x_1/\partial x_i) + (\partial p/\partial x_2)(\partial x_2/\partial x_i) + \ldots \\ &+ (\partial p/\partial x_n)(\partial x_n/\partial x_i)]x_i - \partial C_i/\partial x_i \\ =& p + p'\left[1 + \sum\nolimits_{k\neq i}(\partial x_k/\partial x_i)\right]x_i - c_i' = \{p + p'\,[1+\theta_i]x_i\} - c_i' = 0.^{7}\end{aligned} \tag{2}$$

The term $\theta_i \equiv \Sigma_{k\neq i}\,(\partial x_k/\partial x_i)$ represents the sum of rival's responses to a change in producer i's volume. It is called *conjectural variations*[8] signaling that producer i holds conjectures about his rival's responses. The entire bracketed term is marginal revenue, and the condition says that optimal production is at a level where marginal revenue equals marginal cost.

From (2) we may extract several behavioral types. The *Nash–Cournot* hypothesis is that producer i conjectures that his rivals do not react to his change of volume, i.e., $(\partial x_k/\partial x_i) = 0, k \neq i$, whereby $\theta_i = 0$. Thus,

$$\partial\pi_i/\partial x_i = \{p + p'x_i\} - c_i' = 0. \tag{2.1}$$

[7] For a homogeneous product $\partial p/\partial x_1 = \partial p/\partial x_2 = \ldots = \partial p/\partial x_n = p'$.

[8] Bowley (1924) introduced the idea. Game theorists dislike the concept – only the Nash solution $\theta_i = 0$ is consistent. Others point out that $\theta_i \neq 0$ in particular markets and that should be taken as a datum along with costs and demand. Bresnahan (1989) reviews several studies where such terms have been estimated.

Bertrand's model can be described as $\theta_i = -1$, whereby $[1 + \theta_i] = 0$. This is as if the Bertrand firm believes that its increased output will be exactly offset by the other firms. Observe that we are left with $p - c_i' = 0$. This model should be distinguished though from the case of the *price taker* who ignores his influence on price, i.e., he thinks $p' = 0$, in which case we also are left with the condition that price equals marginal cost,

$$\partial \pi_i / \partial x_i = p - c_i' = 0. \tag{2.2}$$

The most general behavior is the case where $(\partial x_k / \partial x_i)$ can be of any sign and value, whereby $\theta_i \neq 0$ in (2). The *Stackelberg* model is one example and the model of a *dominant firm* with a fringe of price-takers is another.[9]

Consider a *cartel* of producers, $i \in \Lambda \subseteq N = \{1, 2, \ldots, n\}$,[10] who maximize the sum of profits. Each member i of the cartel thus considers the total production of the cartel $(\Sigma_{k \in \Lambda} x_k)$ and not only his own (x_i) when he adjusts his production

$$\partial \pi_i / \partial x_i = \left\{ p + p' \left(\sum\nolimits_{k \in \Lambda} x_k \right) \right\} - c_i' = 0. \tag{2.3}$$

3.1 Non-Negative Sales

In applications one has to allow the possibility that a producer does not operate profitably in the market. That is, the maximization problem of producer i is

$$\textit{maximize}\, \pi_i = \left\{ p(X)\, x_i - C_i(x_i) \right\} \text{ subject to } x_i \geq 0, i = 1, \ldots, n. \tag{3}$$

The first-order (Kuhn–Tucker) conditions of a Nash–Cournot player are:

$$\begin{aligned} &- \partial \pi_i / \partial x_i = -\{p + p' x_i\} + c_i' \geq 0, \\ &x_i \geq 0, \quad \text{and} \\ &x_i \left[\{p + p' x_i\} - c_i' \right] = 0, i = 1, \ldots, n. \end{aligned} \tag{3.1}$$

The first part of (3.1) states that marginal profit on sales has to be nonpositive. Assume the opposite, namely that $\partial \pi_i / \partial x_i$ was positive. Then, one could increase profits by expanding sales, invalidating the initial position as equilibrium. The second condition is that the volume has to be non-negative. Finally, if a volume is positive, the marginal profit is zero, and oppositely, if marginal profit is negative, the volume is zero. It simply is not profitable to sell even the first unit.

[9] A leader assumes that his followers adjust their volumes to satisfy their individual first order conditions. Their aggregate response follows from totally differentiating these conditions. The difference between the models stems from the different behavior of the followers – Cournot-players versus price-takers.

[10] It is assumed that the cartel behaves as one Nash–Cournot player against non-members. Thus, $\theta_i = 0$.

The first-order conditions (3.1) stem from n different optimization problems that are interrelated. Mathiesen (1985) suggested that this problem could be solved by the SLCP-algorithm,[11] which today is available as the MCP-solver in GAMS.[12]

3.2 *Optimization Approach*

A traditionally more popular approach is to try to convert equilibrium conditions like (3.1) into a single optimization problem.[13] The question is, under what conditions does there exist a (fictitious) function $\Pi(\mathbf{x})$ with the property that:

$$\partial\Pi/\partial x_{\mathrm{i}} = \partial\pi_{\mathrm{i}}/\partial x_{\mathrm{i}}, i = 1, \ldots, n. \tag{4}$$

If such a function $\Pi(\mathbf{x})$ exists, the game of n agents maximizing individual profit functions would be equivalent to a problem of a single agent maximizing this fictitious objective, and it could be solved by any optimizing code. In general, the function $\Pi(\mathbf{x})$ exists if and only if the n first-order conditions, like (3.1), are integrable.[14] Slade (1994) showed that for the homogeneous product Cournot model a function $\Pi(\mathbf{x})$ exists when (inverse) demand is linear, i.e., $p = a - bX$. In this case,

$$\Pi\,(\mathbf{x}) = \Sigma_{\mathrm{i}}\pi_{\mathrm{i}} - [1/2\, b\Sigma_{\mathrm{i}} x_{\mathrm{i}}(\Sigma_{\mathrm{j}} x_{\mathrm{j}})]. \tag{4.1}$$

It is well-known that the maximization of aggregate profits (the first term) leads to a collusive outcome. The bracketed term thus corrects for this erroneous objective; it undoes the coordination of activities implied by the sum of profits.

4 The Generalized Transportation Model

Consider a market for a homogeneous good. Assume there are many producers and consumers so that one may reasonably apply price taking behavior. Individual producers' supply is aggregated into supply curves (industry cost curves) per region, and likewise, individual consumers' demand is aggregated into demand functions

[11] The acronym stands for A Sequence of Linear Complementarity Problems, describing a Newton-like iterative process where the linear conditions in each step are solved by Lemke's method.

[12] GAMS is a software package for a variety of optimization and equilibrium problems. It has become an industrial standard. See www.gams.com for details on content and how to obtain this package.

[13] Samuelson (1952) originated this approach, demonstrating that in order to compute the competitive equilibrium one could maximize the sum of consumers' and producers' surpluses. See Takayama and Judge (1971) for applications.

[14] Cf., integrability of demand in economic theory (see, e.g., Varian 1992).

per region or segment. Let there be n producing and m consuming regions, and let i and j denote producing, respectively, consuming region. Further, let

c_i denote the marginal cost of production in region i,
$s_i(c_i)$ denote supply from producing region i,
x_{ij} denote sales from producing region i to consuming region j,
t_{ij} denote the unit transportation cost from i to j,
p_j denote consumer price in consuming region j, and
$d_j(p_j)$ denote demand in consuming region j.

A competitive equilibrium is characterized by three sets of conditions.

Supply balance:

$$s_i(c_i) - \sum\nolimits_j x_{ij} \geq 0,\ c_i \geq 0, c_i\left[s_i(c_i) - \sum\nolimits_j x_{ij}\right] = 0,\ i = 1, \ldots, n. \tag{5.1}$$

Demand balance:

$$\sum\nolimits_i x_{ij} - d_j\left(p_j\right) \geq 0,\ p_j \geq 0,\ p_j\left[\sum\nolimits_i x_{ij} - d_j\left(p_j\right)\right] = 0, j = 1, \ldots, m. \tag{5.2}$$

Price formation:

$$c_i + t_{ij} - p_j \geq 0, x_{ij} \geq 0, x_{ij}\left[c_i + t_{ij} - p_j\right] = 0, i = 1, \ldots, n, j = 1, \ldots, m. \tag{5.3}$$

Equations (5.1) and (5.2) are conditions on regional balances of supply and demand, while (5.3) relates to the profitability of trade flows between regions. In (5.1), the production of region i has to be at least as large as its total sales; marginal cost – interpreted as the supply-price – has to be non-negative; and finally, if the supply price is positive, production equals sales. Equation (5.2) says that consumption in region j cannot be larger than deliveries; the price has to be non-negative, and finally, if the price is positive, consumption equals total delivery. Equation (5.3) parallels condition (3). The negative of marginal profit on sales from supply-region i to consuming region j has to be non-negative. The flow has to be non-negative; and, if a flow is positive, its marginal profit is zero, while on the other hand, if marginal profit is negative, the flow is zero. As we shall see below, zero will be a typical outcome for a large number of flows in a competitive equilibrium.

Consider now maximizing the sum of consumers' and producers' surpluses. Let

$c_i(Q_i) = c_{0i} + c_{1i}Q_i$ denote the marginal cost of producing Q_i in region i, and
$b_j(Z_j) = a_j - b_jZ_j$ denote marginal willingness to pay for consumption Z_j.

The optimization approach to solving (5.1)–(5.3) can then be stated as

$$\textit{maximize}\left\{\sum\nolimits_j\left(a_j - {}^1\!/_2 b_j Z_j\right)Z_j - \left[\left(\sum\nolimits_i c_{0i} + {}^1\!/_2 c_{1i}Q_i\right)Q_i + \sum\nolimits_i\sum\nolimits_j t_{ij}x_{ij}\right]\right\} \tag{6.1}$$

$$\text{subject to } \sum\nolimits_{\mathrm{i}} x_{\mathrm{ij}} \geq Z_{\mathrm{j}}, \sum\nolimits_{\mathrm{j}} x_{\mathrm{ij}} \leq Q_{\mathrm{i}}, x_{\mathrm{ij}} \geq 0. \tag{6.2}$$

Equations (6.1) and (6.2) constitute a quadratic programming problem analyzed by Takayama and Judge (1971). The objective (6.1) represents consumers' and producers' surpluses computed as the difference between consumer valuation of volumes (Z_{j}) and costs of production and transportation. If Q_{i} and Z_{j} are exogenously stipulated production and consumption volumes (Q'_{i} and Z'_{j}), (6.1) and (6.2) reduce to minimizing transportation costs. This is the transportation model of linear programming.

$$\begin{aligned} &\textit{maximize } \left\{-\sum\nolimits_{\mathrm{i}} \sum\nolimits_{\mathrm{j}} t_{\mathrm{ij}} x_{\mathrm{ij}}\right\} \\ &\text{subject to } \sum\nolimits_{\mathrm{i}} x_{\mathrm{ij}} \geq Z'_{\mathrm{j}}, \sum\nolimits_{\mathrm{j}} x_{\mathrm{ij}} \leq Q'_{\mathrm{i}}, \text{ and } x_{\mathrm{ij}} \geq 0. \end{aligned} \tag{6.3}$$

4.1 Noncompetitive Behavior

Assume now that producer i is a single decision unit and not an aggregate of individual producers. Assume further that this agent behaves according to the Nash hypothesis when selling in region j, i.e., he observes his influence on price (p_{j}) and conjectures that other producers regard his volume (x_{ij}) as given. The first-order condition for his profitable sale to region j is then

$$\begin{aligned} &(c_{\mathrm{i}} + t_{\mathrm{ij}}) - (p_{\mathrm{j}} + p_{\mathrm{j}}' x_{\mathrm{ij}}) \geq 0, x_{\mathrm{ij}} \geq 0, \quad \text{and} \\ &x_{\mathrm{ij}} \left[(c_{\mathrm{i}} + t_{\mathrm{ij}}) - (p_{\mathrm{j}} + p_{\mathrm{j}}' x_{\mathrm{ij}})\right] = 0, i = 1, \ldots, n, j = 1, \ldots, m. \end{aligned} \tag{5.3$'$}$$

His marginal profit is now the difference between the marginal revenue and the marginal cost of production plus transportation. This has to be nonpositive.

Let us compare implications of different behaviors. The price-taker sells to markets considering his net-back price $[p_{\mathrm{j}} - t_{\mathrm{ij}}]$, i.e., his decision rule is sell to region j when

$$p_{\mathrm{j}} - t_{\mathrm{ij}} \geq p_{\mathrm{k}} - t_{\mathrm{ik}} \text{ for all } k \neq j. \qquad \text{See(5.3).} \tag{7.1}$$

Knowing that his supply affects the price, the Nash player considers the net-back of marginal revenue $[(p_{\mathrm{j}} + p_{\mathrm{j}}' x_{\mathrm{ij}}) - t_{\mathrm{ij}}]$, i.e., his decision rule is sell to region j when

$$\left[\left(p_{\mathrm{j}} + p_{\mathrm{j}}' x_{\mathrm{ij}}\right) - t_{\mathrm{ij}}\right] \geq [(p_{\mathrm{k}} + p_{\mathrm{k}}' x_{\mathrm{ik}}) - t_{\mathrm{ik}}] \text{ for all } k \neq j. \tag{7.2}$$

This net-back is an explicit function of his volume (with $p' < 0$), and he adjusts x_{ij} so that sale to market j may be profitable. Hence, for a market far away, where net-back price will be low, he will only sell a little. The following example will illustrate the implications for trade volumes of rules (7.1) versus (7.2).

4.1.1 A Numerical Example

Consider a market consisting of *four* individual producers and *six* consuming regions. For the illustration of consequences of different behavioral assumptions we use identical linear (inverse) demand and identical linear marginal cost functions

$$p_j = 20 - 0.25 Z_j,\ j = 1, \ldots, 6, \text{and}\, c_i = 1 + 0.1\, Q_i,\ i = \text{A, B, C, D}.$$

Unit transportation costs differ and are shown in Table 1.

For this model we have computed the competitive and the Cournot equilibria. They differ in several respects. At the aggregate level, the competitive equilibrium has a larger volume (262 vs. 214) and a lower (average) price (9 vs. 11). This is well known. See Tables 2 and 3. Even though demand and marginal cost functions are identical across regions, unequal transportation costs make regional consumption and individual production volumes differ. These are also well known facts.

Volumes and prices vary much more (in both absolute and relative terms) across regions in the competitive equilibrium than in the Cournot equilibrium. The largest volume is 26% and the highest price 32% above the smallest volume (price) in the competitive equilibrium. The corresponding numbers in the Cournot equilibrium are 8% for both volume and price. Price variation between regions in a competitive equilibrium is bounded by differences in transportation costs. It is interesting to observe

Table 1 Unit transportation cost

From/to	1	2	3	4	5	6
A	1.5	2.5	2	3	4	3.5
B	0.5	1.6	0.99	1.5	3	2.5
C	3.5	3.5	2	1	2.5	4
D	4	3.5	3.5	2.5	1	2

Table 2 Volumes and prices of the competitive equilibrium

From/to	1	2	3	4	5	6	Sum	mc
A	17.0	42.3					59.3	6.9
B	29.3		40.0				69.3	7.9
C			4.4	48.3	6.5		59.2	6.9
D					35.8	38.3	74.2	8.4
Sum	46.3	42.3	44.3	48.3	42.3	38.3	261.9	
Price	8.4	9.4	8.9	7.9	9.4	10.4	9.0	

Table 3 Volumes and prices of the Nash–Cournot equilibrium

From/to	1	2	3	4	5	6	Sum	mc
A	13.3	10.5	10.5	6.1	4.1	7.3	51.7	6.2
B	14.3	11.2	11.5	9.1	5.1	8.0	59.1	6.9
C	5.3	6.5	10.5	14.1	10.1	5.3	51.7	6.2
D	3.3	6.5	4.5	8.1	16.1	13.3	51.7	6.2
Sum	36.1	34.8	36.9	37.3	35.3	34.0	214.2	
Price	11.0	11.3	10.8	10.7	11.2	11.5	11.1	

that price differences in the Cournot equilibrium may be significantly smaller. Of course, with identical demand, price elasticities are equal between regions in this example, whereby a Cournot producer has no incentive to price discriminate.[15] In reality, price elasticities differ between various consumer groups. In a gas market context, the relevant questions are whether a producer can price discriminate between consumers[16] and to what extent the mix of consumer groups (with different price elasticities) vary between regions.

The more spectacular difference between these equilibria is the rather diverse trade patterns, with nine positive flows in the competitive equilibrium and 24 in the Cournot equilibrium.[17] In general and unless unit transportation costs (t_{ij}) of producer i differ too much between regions, a Cournot producer will supply all regions. The price taking producer, however, supplies only a few of his neighboring regions. The rationale of the competitive equilibrium is to provide commodities at the lowest cost.[18] A competitive equilibrium would for example never have Norway supplying Italy or Spain, *and* at the same time Algeria supply Belgium or Germany. Such a solution would imply hauling a homogeneous commodity both ways across the Alps and would constitute a waste of resources.[19] This trade pattern, however, characterizes the Cournot equilibrium,[20] and to some extent also the present gas market.[21]

From consumers' perspective of security of deliveries it is noteworthy that the Cournot model provides a dramatically different portfolio of suppliers. All suppliers sell in every region and the market share of the dominant supplier varies between 0.26 and 0.38. This is a desirable property, although it may be unintended. In comparison, the competitive equilibrium has one or at most two suppliers and market shares vary between 0.63 and 1.[22]

---

[15] With $(p, Z) = (10, 40)$, price elasticity is -1. When elasticities differ between regions (-0.6 to -1.4), price differences in the Cournot equilibrium are still smaller.

[16] A gas producer typically does not sell to final consumers, but to various agents who sell and distribute gas to final consumers. A model with successive market power would be more appropriate to capture price discrimination (cf., Boots et al. 2004 and Eldegard et al. 2001).

[17] The Cournot equilibrium may have all nm flows positive, while the competitive equilibrium has at most $(n + m - 1)$ positive flows. The latter fact is well known within operations research as a feature of a *basic solution* to the transportation model and within trade theory through the notion of *no cross-hauling*.

[18] It is interesting to note that a collusive outcome – where all producers coordinate their sales (see 2.3) – has the same efficient trade pattern as the competitive solution, although volumes are about 40% lower.

[19] Of course, even though contracts involve sales from North to South and vice versa, transmission companies may avoid actual cross-hauling.

[20] In reality, fixed costs of deliveries block small flows. In modeling terms, this feature implies non-convexities that are hard, but not impossible, to implement in an optimization or equilibrium model.

[21] In line with the regional disaggregation of this example, the European natural gas market has at present about 20 positive flows. When asked why Statoil would sell to Italy rather than Germany, a sales representative commented: "We already sell much in Germany and selling larger volumes would depress our prices there."

[22] Of course, deliveries from any supplier could be constrained in the competitive model in order to prevent such dominance.

A producer's incentive for selling differs between the two models. From (7.1) it follows that he sells only to markets yielding the highest net-back price. Producer A, for example, has net-back prices 6.93, 6.93, 6.92, 4.92, 5.42, and 6.92 from markets 1 to 6. He sells to markets 1 and 2, and does not sell in markets 3 and 6, even though his net-back prices from these two markets are only slightly lower, namely 6.92 versus 6.93. In the Cournot equilibrium, net-back *marginal revenues* are equalized, while net-back *prices* may differ between the markets he supplies (see (7.2)). Producer A sells to all six markets and his net-back prices are 9.5, 8.8, 8.8, 7.7, 7.2, and 8. It is noteworthy that he sells to markets with a net-back price difference of 2.3, while the price taker does not sell to markets where the difference is as low as 0.01.[23]

The regional price-pattern of the two models also follows different logics. An immediate result of the competitive equilibrium is that when a supplier sells in several markets, the price increases with transportation cost (distance to the market) and covers his combined costs of production and transportation. The Nash–Cournot equilibrium, however, may have decreasing price with transportation cost (distance) to a market as illustrated in the following stylized example. Consider three regions 1–3 along a line. There are pipelines with a unit transportation cost of 1 between neighboring markets. Gas suppliers are located in regions 1 and 3, but not in region 2. Demand in a region is $Q = 15 - P$, and the industry cost curves of regions 1 and 3 are, respectively,

$$c_1 = 3 + 0.25\, Q_1 \text{ and } c_3 = 8 + 0.4 Q_1.$$

In a competitive equilibrium producers of region 1 sell in all markets, while producers of region 3 only sell in their home market. Equilibrium prices are 7, 8, and 9, respectively.

Assume now that there is one supplier in region 1 with the marginal cost curve c_1 shown above and two suppliers in region 3 each with a marginal cost curve $c_{3k} = 8 + 0.2 q_k$, $k = 1, 2$, which may be interpreted as each having half the supply of the aggregate (Q_1). The Nash–Cournot equilibrium has prices 10.33, 10.28, and 10.03, i.e., decreasing and not increasing going from 1 to 3.[24]

[23] Traded volumes in the competitive equilibrium depend on transportation cost in a step-like manner, that is, a volume stays fixed for cost variations over some range and then jumps. The difference in net-back prices between a market that is served and one that is not served may be arbitrarily small. In our example, N does not serve market 3 until transportation cost is below 1.99. But at a unit cost of 1.98 A shifts his entire volume (17) from market 1 to market 3 and B shifts a similar volume oppositely. Who sells where in a competitive equilibrium is therefore very sensitive to relative transportation costs. Individual flows in the Cournot equilibrium are much more stable to such parameter changes. (Aggregate flows of production and consumption, however, are very stable in both models.)

[24] When establishing a gas model for the European market in the mid-1980s it seemed that actual gas-prices were reduced going east to west in Europe. The reason could be that USSR met no competition in eastern markets, while there were alternative suppliers and hence competition in the West.

Based on the observation that price exceeds supply cost in the Cournot equilibrium, it has been suggested (Smeers 1997) that by allowing mark-ups on cost, the competitive model can be employed to simulate the Cournot equilibrium. Inspection of conditions (5.3) and (5.3′) reveal that a mark-up-factor (λ_{ij}) has to be

$$\lambda_{ij} = -(p_j' x_{ij})/(c_i + t_{ij}), i = 1, \ldots, n, \text{and } j = 1, \ldots, m.$$

There are two problems here. One is that the information content is demanding as these mark-ups differ both by producer and market. In fact, the entire Cournot flow-matrix (x_{ij}) has to be known. So, if that is known, why bother use another model to replicate it?

The next problem is related to the model, its solver and the structure of the solution. Assume the Cournot equilibrium is known and compute parameters λ_{ij}, as distinguished from functions of variables. Solve (6.1) and (6.2) modified by mark-ups $(1 + \lambda_{ij})$ on cost. The resulting model has multiple solutions. One is the Cournot equilibrium another has the trade pattern of Table 2 only flows are smaller. But there are more solutions. With multiple solutions, it is my experience that it is unclear which solution will be reported by the solver.[25] It is likely to be a competitive-looking one. As illustrated above, such a solution may be inappropriate for some managerial purposes, e.g., analyzing marketing strategies, because its trade pattern is so sensitive to (direct or indirect) changes in transportation costs.

5 Network

Transportation in the previous model can be pictured as in Fig. 2a. Supplier i can sell in any market j. There is no explicit mention of how transportation is performed, and only unit transportation cost is represented. It is assumed to be the lowest cost of shipping a unit from i to j. This is a *transportation type* model.

Natural gas has to follow a designated transportation system, a network of pipelines, although the gas may flow along any of a number of different routes. Figure 2b displays the same industry where producers (A–D) insert gas into the network and consuming regions (1–6) extract gas. Gas from A to market 6, for example, may flow along one or more of several routes: A-2–4–6, A-1–4–6, A-1–3–4–6, or A-1–3–5–6. This a *network type* of model.

[25] The MCP-solver of GAMS replicated the Nash equilibrium when the mark-up was computed immediately following the computation of the (true) Nash, but it generated a trade pattern with 9 flows when the mark-up equilibrium was computed immediately following the competitive one (or when the matrix of λ_{ij}'s was input to a new run). Although both these trade patterns had 9 positive flows only two flows were positive in both equilibria. The reason for these seemingly arbitrary trade patterns in the mark-up equilibrium stems from the definition of λ_{ij} whereby reduced costs $\{p_j - (c_i + t_{ij})(1 + \lambda_{ij})\} = 0$, all i, j, at the equilibrium admitting any trade pattern that adds up to equilibrium production and consumption volumes.

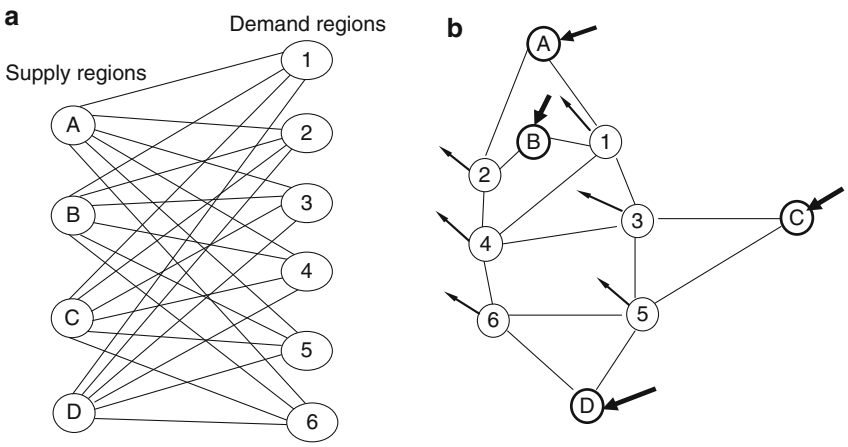


Fig. 2 (**a**) Transportation type. (**b**) Network type

From a network model one may compute the cheapest route of transporting between two regions, and use the corresponding unit cost in a transportation type model. The critical issue is whether there are features of the network that are not represented by such unit costs.

Restricted capacities of individual pipelines are certainly not represented in the simplest version of the transportation model. Figure 3a illustrates the case of a pipeline with unit cost, or tariff, t and a given maximal capacity K (in terms of m^3 flow per time period). As long as the total flow along this pipeline is below K, e.g., demand for capacity is D_1, there is no problem. All shipments are allowed at the unit cost t. If demand for capacity exceeds K, like in case D_2, however, some shipments cannot be executed, and there has to be rationing. The simplest rationing from a modeling point of view is to use the shadow price μ of the scarce capacity K. That is, the cost per unit transported through the pipeline is increased to $(t + \mu)$, at which demand equals the capacity. Assume that the value of μ is known. Tariff t could then be replaced by $(t + \mu)$ and the capacity constraint disregarded. The aggregate flow in the optimal solution would now equal the capacity.

The shadow price depends on the optimal solution of the market model, however, much like the mark-ups of the previous section. Hence, shadow prices cannot be known in advance, although the value for a particular pipeline may turn out to be fairly stable between various scenarios.

A critical issue is whether the time horizon of the analysis is sufficiently long so that capacities can be increased in order to be nonbinding. Then one may assume that the necessary investments will be made to accommodate planned flows, and one may employ a transportation type model. If, on the other hand some capacities will be constraining within the time horizon, there may be no escape, but model the actual network. Obviously, one will have to aggregate details also in this dimension of the market and represent only the major transmission lines.

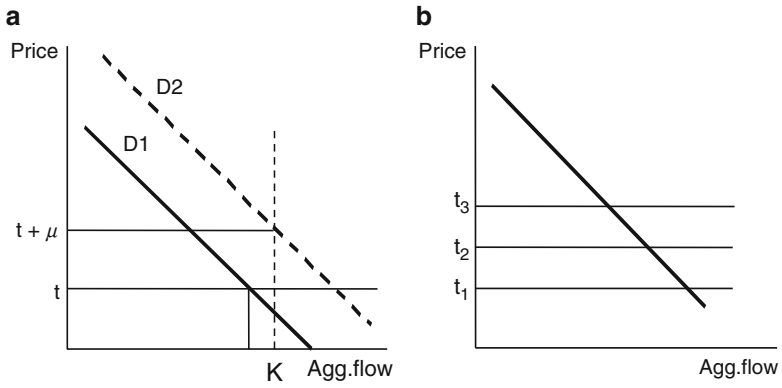


Fig. 3 (**a**) Pipeline pricing. (**b**) Pipeline pricing

There are two issues related to behavioral types when modeling a network of individual pipelines. One is about gas suppliers the other is about pipeline owners. The first is about how the flows are modeled. Traditionally, flows into and out of a given node are aggregated, and the aggregate flow into, e.g., node 4 (in Fig. 2b) has to equal the aggregate flow out of this node. The model sums the cost per unit of gas injected into the network and the unit costs of transportation (including shadow prices) along the various pipelines into a price of delivering gas at any node. And oppositely, it provides net-back prices to a supplier from any market.

In an application with several suppliers injecting flows possibly at different nodes, these flows mix and the model disregards who are the sellers. This is of little importance if all sellers are price takers (see (5.3)). The information on market prices is available and the model computes net-back prices of supplying the various markets. When supplier i is not a price-taker, however, for example he may be a Nash–Cournot player, information on his supplies (x_{ij}) to market j is needed (see (5.3′)). In order to provide this information, the model has to keep track of individual sales, which requires a reformulation of the traditional network models.

In the discussion related to Fig. 3a, we stated that at a tariff (or price) $(t + \mu)$, the market involving supply of and demand for capacity of the pipeline was cleared. Implicitly, we assumed that the owner of the pipeline stipulated his price t independent of actual demand or stipulated $(t + \mu)$ taking marginal willingness to pay for this capacity as given. The price taking assumption may be dubious when for example a pipeline owner is in a position to set his price optimally. As mentioned above noncooperative theory does not provide a unique solution when both sides of a market enjoy market power, e.g., a supplier of transmission capacity and a gas producer demanding such capacity. Computing the equilibrium for alternative levels (t_1, t_2, t_3) of the tariff provide corresponding levels of aggregate flows whereby one finds a pipeline owner's corresponding profit and may infer his optimal price (see Fig. 3b).

6 Conclusion

Our concern has been to communicate how to model in order to gain insight into the market mechanism and price formation in the European market for natural gas. When insight and not predictions is the driving force of modeling, we advocate a balance between different features and the amount of factual detail in the description of any such feature. One should appreciate the insights gained from a combination of behavior and geography (Mathiesen 1988), or behavior and time (Brekke et al. 1987), both of which are not available in a very detailed optimization model like, e.g., the McKinsey model (Øygard and Tryggestad 2001).

As exemplified above, the competitive and the Cournot equilibria differ markedly. When analyzing sales strategies at the aggregate level, e.g., selling in Europe, or at a disaggregate level, e.g., selling in France, understanding the competition and the market mechanisms is essential. And in order to understand price formation one has to include rivalry.

Acknowledgment This paper summarizes my experience of gas market modeling. Very helpful comments from two referees are acknowledged.

References

Beltramo, M. A., Manne, A. S., & Weyant, J. P. (1986). A North American gas trade model (GTM). *The Energy Journal, 7*(5), 15–32

Berg, M. (1990). *Økt tilbud av naturgass fra Canada til USA: En markedsstrategisk analyse. (Increased supply of natural gas from Canada to the United States: A strategic market analysis). Working paper 16/90*. Bergen: Senter for Anvendt Forskning (Center for Applied Research)

Bjerkholt, O., Gjelsvik, E., & Olsen, Ø. (1989). *Gas trade and demand in Northwest Europe: Regulation, bargaining and competition*. Discussion Paper 45/89. Oslo: Statistics Norway

Boots, M. G., Rijkers, F. A. M., & Hobbs, B. F. (2004). Trading in the downstream European gas market: A successive oligopoly approach. *The Energy Journal, 25*(5), 73–102

Boucher, J., & Smeers, Y. (1984). *Simulation of the European gas market up to the year 2000*. Discussion paper 8448. Louvain-la-Neuve: CORE

Bowley, A. L. (1924). *The mathematical groundwork of economics*. Oxford: Oxford University Press

Brekke, K. A., Gjelsvik, E., & Vatne, B. H. (1987). *A dynamic supply side game applied to the European gas market*. Discussion paper 22/87. Oslo: Statistics Norway

Bresnahan, T. F. (1989). Studies of industries with market power. In R. Schmalensee & R. D. Willig (Eds.), *The handbook of industrial organization*. Amsterdam: North Holland Publishing Company

De Wolf, D., & Smeers, Y. (1993). *Optimal dimensioning of pipe networks with an application to gas transmission networks*. Discussion paper 9315. Louvain-la-Neuve: CORE

Eldegard, T., Mathiesen, L., & Skaar, J. (2001). *A partial equilibrium model of the European natural gas market*. SNF Report 52/01. Bergen: Stiftelsen for samfunns- og næringslivsforskning (Institute for Research in Economics and Business Administration)

Fuglseth, A. M., & Grønhaug, K. (2003). Can computerised market models improve strategic decision-making? An exploratory study. *Journal of Socio-Economics, 32*(5), 503–520

Golombek, R., Gjelsvik, E., & Rosendahl, K. E. (1995). Effects of liberalizing the natural gas markets in Western Europe. *The Energy Journal, 16*(1), 85–112

Golombek, R., Gjelsvik, E., & Rosendahl, K. E. (1998). Increased competition on the supply side of the Western European natural gas market. *The Energy Journal, 19*(3), 1–18

Grais, W., & Zheng, K. (1996). Strategic interdependence in European East–West gas trade: A hierarchical Stackelberg game approach. *The Energy Journal, 17*(3), 61–84

Hoel, M. et al. (1987). *The market for natural gas in Europe: The core of a game*. Memo 13. Oslo: Department of Economics, University of Oslo

Keller, L. (2001). BP. Presentation at NPF Trondheim. May 28, 2001

Mathiesen, L. (1985). Computational experience in solving equilibrium models by a sequence of linear complementarity problems. *Operations Research, 33*, 1225–1250

Mathiesen, L. (1987). *GAS: En modell og et interaktivt system for analyser av det vest-europeiske marked for naturgass (A model and an interactive system for analysis of the Western European market for natural gas)*. Report 3. Bergen: Senter for Anvendt Forskning (Center for Applied Research)

Mathiesen, L. (1988). *Analyzing the European market for natural gas*. Working paper 37/88. Bergen: Senter for Anvendt Forskning (Center for Applied Research)

Mathiesen, L., Roland, K., & Thonstad, K. (1987). The European natural gas market: Degrees of market power on the selling side. In R. Golombek, M. Hoel, J. Vislie (Eds.), *Natural gas markets and contracts*. Amsterdam: North Holland Publishing Company

Samuelson, P. (1952). Spatial price equilibrium and linear programming. *American Economic Review, 32*(4), 510–524

Slade, M. (1994). What does an oligopoly maximize? *Journal of Industrial Economics, 57*(2), 45–62

Smeers, Y. (1997). Computable equilibrium models and the restructuring of the European electricity and gas markets. *The Energy Journal, 18*(4), 1–31

Takayama, T., & Judge, G. G. (1971). *Spatial and temporal price allocation models*. Amsterdam: North Holland Publishing Company

Tandberg, E. (1990). *Naturgass i det nordiske energimarkedet (Natural gas in the Nordic energy market)*. Rapport 4/90. Bergen: Senter for Anvendt Forskning (Center for applied research)

Tandberg, E. (1992). *Introduction of natural gas in the Scandinavian energy market*. HAS-thesis NHH and working paper 53/92. Bergen: Stiftelsen for samfunns- og næringslivsforskning (Institute for Research in Economics and Business Administration)

Varian, H. (1992). *Microeconomic analysis*. New York: Norton and Co

von Hirschhausen, C., Meinhart, B., & Pavel, F. (2005). Transporting Russian gas to Western Europe: A simulation analysis. *The Energy Journal, 26*(2), 49–68

Øygard, S. H., & Tryggestad, J. C. (2001). Volatile gas. *McKinsey Quarterly, 3*

Equilibrium Models and Managerial Team Learning

Anna Mette Fuglseth and Kjell Grønhaug

Abstract The purpose of this paper is to enhance the insights of whether and, if so, how equilibrium models can enhance managerial team learning in complex and ever-changing situations. To handle this research challenge we first clarify the concept of managerial team learning. Then, we present an example of an equilibrium model, VisualGas, which has been developed to help managers in a large oil and gas company to improve their understanding of market behavior. We report on a quasi-experimental study where a team of managers/market analysts evaluated the consequences for the company of a critical market event, first without and then with the system. It was found that use of the system led to a more varied interpretation of the event, i.e., learning, but also to a focus on the system variables only.

1 Introduction

The purpose of this paper is to explore whether equilibrium models can enhance managerial team learning in complex and ever-changing environments – and if so, how?

In the past decades, there has been an increasing focus on the ability of organizations to learn. There are several reasons for this increased focus on learning, such as rapid technological development and political changes, which also have led to liberalization and globalization of markets, and consequently intensified competition. Furthermore, these developments have contributed to increased turbulence in firms' environments, and thus their needs to discover opportunities and threats to survive and prosper. To handle such changes adequate knowledge is needed, and thus continuous learning becomes a prerequisite. It is the individual organization member who learns. Today, organizations are usually headed by a team of managers. This implies that the learning in the managerial team becomes crucial.

A.M. Fuglseth (✉) and K. Grønhaug
Department of Strategy and Management, Norwegian School of Economics and Business Administration (NHH), Breiviken 40, 4045 Bergen, Norway
e-mail: anna.mette.fuglseth@nhh.no; kjell.gronhaug@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_7,
© Springer-Verlag Berlin Heidelberg 2010

Review of the literature on managerial team learning reveals, however, that scarce attention is paid to the role of model-based systems in general and in particular to the role of equilibrium models, to support such learning. This is especially the case for important, nonrepetitive strategic tasks, i.e., in situations where the need for learning and knowledge is the greatest.

In economics an equilibrium model is a simplified representation of a single market (partial model) or an economy with interrelationships among various markets (general model). This paper focuses on equilibrium models that are specified according to a mathematical format amenable to interpretation by algorithms for computing equilibrium prices and volumes (see, e.g., Mathiesen 1985). The output from the algorithm is a state of the modeled market or economy in which supply and demand for every good and service are equal. Equilibrium is usually considered to be a hypothetical state, but the market/economy may be thought of as moving towards equilibrium. Equilibrium models may thus be a useful tool to enhance the understanding of consequences of complex market changes.

In this paper, we draw attention to a specific equilibrium model developed to help managers in a large oil and gas company improve their understanding of European gas markets, i.e., turbulent markets of great strategic importance for the company.

The remaining part of this chapter is organized as follows. In the next section, we clarify the notion of managerial team learning. There we also argue for our emphasis on teams. Next, we describe the actual equilibrium model, VisualGas. After this we describe the design of a quasi experiment where a managerial team evaluates the consequences of a critical market event for the company, first without and then with VisualGas. The findings from the two quasi-experimental sessions are compared and related to criteria expected to reflect learning. Implications of the findings are discussed.

2 Managerial Team Learning

Learning as such is an old topic to study. Lately, managerial team learning has increasingly attracted interest both among researchers and practitioners. However, not only is the term (concept) ambiguous, but also the phenomenon itself, including derived questions such as who – or what – are the learners, and where and how does such learning take place (Easterby-Smith et al. 1998).

It is not our purpose here to review the extant and diverse literature related to this topic. Rather, our purpose is to address the research question whether and possibly how equilibrium models may enhance managerial team learning in complex tasks under uncertainty, a topic that has almost been neglected in previous research.

The term managerial team learning implies that it should denote learning beyond individual learning in organizational contexts. In this paper, we have found it useful to build on Huber's (1991) characterization of organizational learning, which also includes team learning:

- Learning occurs when any of the members acquires knowledge that may be related to the organization;
- More learning occurs when the knowledge is transferred to other members;
- More learning occurs when more varied interpretations are developed; and
- When more members comprehend and contribute to such varied interpretations.

In this paper, we are particularly interested in assessing learning as development of "more varied interpretations" in a managerial team. The reason is that interpretation, i.e., explaining, construing the significance of events, is an essential element of organizational performance (Drucker 1974). Real-world complex events do not present themselves as objective facts. Events must be noticed and interpreted, often based on weak signals and data from various sources.

Our reasons for building on Huber's view are that it does not limit organizational learning to a process initiated by a mismatch between expected and actual results of action, but also include learning as a process initiated to respond to plausible future events (Nonaka 1994), which is essential for our research purpose.

One limitation of the above characterization is that it does not relate learning to the history of the organization, i.e., that learning depends on what is already known in the organization (Argyris and Schön 1996). Another limitation is that it does not link individual and group/team learning to organizational references that are established to guide behaviors (Argyris and Schön 1996), e.g., goals, strategies, policies and routines. Such linking is necessary to understand how individual and group/team learning can lead to concerted activities that increase organizational effectiveness.

In contrast to Huber (1991), we use the notion "member" instead of terms such as "unit," "component," and "entity." The reason is our belief that learning in teams is to be understood as both a cognitive and a social activity (Gherardi et al. 1998). Our contribution is based on the view that it is individuals who learn (Simon 1991; Argyris and Schön 1996), but also that individuals are social beings who construct their understanding and learn from social interaction, among others in the workplace (Gherardi et al. 1998).

Management literature increasingly emphasizes the importance of teams for organizational success in the modern economy (Cohen and Bailey 1997; Senge 1990). A team is a small number of people with complementary competences and skills, committed to a common purpose (Katzenbach 1998). There are multiple reasons for the increased emphasis on teams. The handling of complex tasks often requires complementary competences and skills that are not possessed by one individual only. Furthermore, in case of environmental changes, a variety of perspectives are often useful to enhance the understanding of causes and possible consequences of events, which is again essential for the development of adequate actions.

The concept of learning is usually related to individual living organisms' adaptation to their particular environment (Anderson 1995). A crude distinction can be made between theories of behavioral and cognitive learning (Anderson 1995). Behavioral theories attempt to explain learning as a result of training or reactions to performance feedback without considering conscious thought. Cognitive learning theories attempt to explain variations in performance by considering changes in individuals' knowledge structures.

In this paper, we focus on cognitive learning, i.e., learning related to generating and acquiring new knowledge, new ways of perceiving problems and generating solutions, etc. Such learning requires knowledge and memories for storing knowledge and representations of knowledge, and it requires processes to change/develop the knowledge.

Distinctions are often made between different phases of the learning process (see Pawlowsky 2001). Huber (1991), e.g., distinguishes among knowledge acquisition, information distribution, information interpretation and memorization. In our opinion the basic processes as regards team learning are information and communication processes. We believe that focus on these processes helps understand how knowledge is developed and transferred among individuals with different knowledge structures. Information processing comprises the individual's detection of data and other stimuli from the environment, interpretation of the data/stimuli, reflection and the coding of information as data to be communicated to others or to be entered into a database. It should be noted, that reflection is not dependent on external stimuli, but can also be a further processing of present knowledge. Communication involves at least two persons, some kind of message and a medium for transfer of messages between the persons. Communication processes are particularly related to interaction involving language, and communication is said to be effective if a receiver understands a message as intended by the sender. Knowledge transfer and development of "more varied interpretations" are in our opinion closely related to the notion of effective communication.

Huber (1991) does not elaborate on the notion of "more varied interpretations." In our opinion, however, such development can usefully be related to a change in the level of information processing at the individual level and group development at the team/group level. Evaluating whether use of equilibrium models support development of more varied interpretations, we will build on cognitive complexity theory (Schroder et al. 1967). This theory focuses on the relationship between the development of knowledge structures and level of information processing, but also includes group development in face-to-face interactions. According to the theory, persons with well developed knowledge structures can function more effectively in complex and changing environments. The theory also argues that level of information processing is influenced by the complexity of the task. If the handling of a complex task makes a heavy demand on an individual's cognitive capacity ("information overload"), the level of information processing may be reduced.

A low level of information processing is characterized by the generation of few alternative interpretations of a stimulus and extensive use of simplifying heuristics, such as availability and anchoring (Tversky and Kahneman 1974) and search for confirmation (see, e.g., Einhorn and Hogarth 1978). Availability refers to the tendency to focus on information that is most "available" in memory and neglect other aspects. Anchoring refers to the tendency to let past focus influence subsequent evaluations. By search for confirmation is meant the tendency to search for data/information that supports prior beliefs and neglect search for data/information that allows to test and challenge prior beliefs and evaluations.

Individuals able to function at a high level of information processing are supposed to be more sensitive to environmental changes and to have an increased perception of uncertainty. They are supposed to take more aspects (dimensions) into consideration when evaluating an event, and to generate more alternative interpretations of the potential effects of the event. In addition, they are supposed to develop more complex interpretations comparing and combining various alternatives, e.g., to view an event from both a positive and a negative perspective and to see an event from the viewpoints of different actors (Schroder et al. 1967; Fuglseth 1989).

As regards group functioning (Schroder et al. 1967, pp. 11–12), it is assumed that group performance basically depends on the group members' level of information processing and on the specific environmental conditions, e.g., task complexity. It is also assumed that the members, in the communication processes, have an effect on the conceptual level of each other, and that group functioning is developmental. Low and high levels of group development are expected to have similar characteristics as described for levels of information processing above.

3 VisualGas

VisualGas is an example of an equilibrium model that has been developed to support strategic decision making and cognitive learning in a Norwegian oil and gas company. The system is based on a generic gas model originally developed by Mathiesen (1987). The model depicts supply and demand functions for the European gas markets. In addition, it handles various forms of producer behavior (price taker, Nash–Cournot oligopoly, and monopoly), pipeline capacity restrictions and tariffs, and import quota. It is connected to an algorithm for solving nonlinear complementarity problems (Mathiesen 1985), i.e., the algorithm calculates equilibria prices and volumes. A graphical user interface takes care of the interaction between the users and the system. The interface presents the gas model to the users, indicates *what* the system can do for them, and *how* they run the system.

The system is primarily used in connection with the strategic decision processes in the company. Experienced managers and analysts are in charge of scenario generation and analyses. Novices participate in the processes and gradually become experienced. When experienced users are transferred to other positions in the organization or leave the company, part of their knowledge is still retained in the strategic analysis group through the discussions with their colleagues and the representation in the system. VisualGas can thus be considered part of the organizational memory (Walsh and Ungson 1991). It is a representation of managers' and analysts' causal beliefs that have developed in information and communication processes in the strategic management group over a long period and with the involvement of many people. It is a tool that provides a point of focus around which to organize the interpretation of market events in the group (Wenger 1998).

4 Research Methodology

Designing an experimental study for real-life strategic situations involves special problems and considerations. First, managers' time is a scarce and expensive resource, so there are usually constraints on the time allocated to participation in experimental studies. Second, managers may be reluctant to participate in research experiments using real-life strategic cases because they do not want to reveal their considerations to researchers who want to publish their findings. Furthermore, there are usually few users of custom-built systems such as VisualGas. It is, therefore, difficult to design an experiment that involves comparison of similar events and a control group.

Our research design is influenced by the above considerations. We had to design a one-group quasi-experimental study. The group consisted of (all) the three users at that time: a manager and two market analysts. The manager had several years of experience using VisualGas for market analyses, and he participated in the latest updating of the model. At the time of the quasi experiment (June 2001), he had, however, not used the system recently. The two analysts were involved in the search and evaluation of the data used for the latest calibration of the model. They are both economists and were well acquainted with the modeling format, but had little experience using the model for analysis of strategic problem situations.

As "experimental stimulus" we chose a case related to the liberalization of the gas markets according to the EU Gas Directive in 1998 and the recent EU critique of the organization of Norwegian gas sales through the Gas Negotiation Committee. The task presented to the group was to evaluate the consequences of the liberalization for the oil and gas company in terms of volumes, prices and profit – if the company does not develop actions to handle the event. The idea for this issue came up in a discussion with one of the top managers. Thus, the task is related to an actual event that had already been noticed as potentially critical. The task was limited to interpretation, i.e., the group was not asked to develop actions also. As argued above, however, interpretation is an essential element of organizational performance.

The quasi experiment included two 2-h sessions with a lengthy lunch in between. First, the subjects analyzed the problem without VisualGas. After lunch, they analyzed the same problem using VisualGas. In the first session, they were allowed to use other information sources and decision aids, such as a spreadsheet.

Data was collected using observation and tape recording of both sessions. The second session was also videotaped. The purpose of the video recording was to synchronize the discussions and the use of VisualGas, i.e., to keep track of which of the screen pages of the system they were commenting on. Furthermore, the scenarios generated during the second session were saved, and we had a copy of the files for analytical purposes. The tape recordings were transcribed.

5 Findings and Analysis

This section presents our findings from the quasi experiment. As argued above, we were particularly interested in evaluating whether the use of VisualGas would lead to development of "more varied interpretations" (Huber 1991) of the market event. Based on our elaboration on this notion, drawing particularly on cognitive complexity theory (Schroder et al. 1967), we derived the following criteria for comparison of the two experimental sessions:

- Number of relevant aspects considered;
- Level of group development as revealed in group discussions (the evaluation criteria are further specified in Table 2 below).

First, we focus on the concepts used by the subjects in each of the two sessions. Then, we give a brief description of each session. Finally, we compare the sessions and relate the findings to the above-mentioned criteria for team learning.

5.1 Concepts Used in the Two Sessions

Table 1 presents concept categories used in the discussions during the two experimental sessions. The purpose of the table is to get an overview of the main differences in concept use without and with VisualGas.

The first column shows how we have categorized the concepts in super ordinate and subordinate concepts. The subordinate concepts are indented. Actual concepts mentioned by the participants are not included in order to save space. In the discussion below, however, we will give examples of the concepts used. The categories have been derived mainly from the structure of the discussions. In the first session, e.g., the participants started with a discussion of the management of the Norwegian shelf with goals and the role of the authorities (see the first category in Table 1). The second column shows the number of concepts used in the discussions during the sessions.

The table reveals considerable differences in concept use between the sessions. One of the main differences is related to the goal discussion. In the first session there was a general discussion of the goals for the Norwegian shelf, such as optimal utilization of resources versus self-interest of the actors. In the second session, focus was on profit generation in the company with consideration of income and cost.

Another difference is that in the first session consequences of the liberalization of the gas markets were almost only related to the supply side, whereas in the second session the consequences were also related to demand. However, the discussion of individual producers was more detailed in the first session, including concepts such as surplus and scarcity of gas, seasonal variation and take-back clauses. In the second session, focus was on the producers' production capacities and volumes.

A third difference is related to categories such as trade pattern and pipelines that were not used in the first session. These categories are closely related to the

Table 1 Concept categories

Concept categories	Number of concepts used in discussions	
	Before	After
Management of the Norwegian shelf	7	1
Goals	2	0
Role of authorities	5	1
Producer economic results	2	7
Profit	1	1
Prices	1	1
Income	0	2
Costs	0	3
Supply (production)	17	13
Producer behavior	4	3
Actors	8	7
Gas fields	0	1
Volumes	5	2
Demand (consumption)	7	16
Actors	3	13
Import quota	0	1
Growth	2	1
Substitutes	2	0
Volumes	0	1
Trade pattern – changes	0	3
Pipe lines	0	4
Number of superordinate categories	4	6
Number of concepts	33	44

consideration of the demand side and to the discussions on profit generation in the second session.

VisualGas did not introduce the participants to new concepts, so the main reason for the differences in concept use is probably the absence of the system to support information processing in the first session. The table indicates that it is particularly difficult to use concepts requiring integration of categories without the support of decision aids. An example is trade pattern that requires integration of demand, supply and pipeline capacity.

The presentation above concerns *which* concepts the team members used in each session, but it does not describe *how* the concepts were used. In the next sections we will further elaborate on the differences in the discussions.

5.2 Description of the First Session (without VisualGas)

The first session started with a discussion of assumptions on which to base the development of consequences of the liberalization of gas markets. The members

assumed that the company would retain some market power, and that the Norwegian authorities would still control the exploitation of the Norwegian gas fields. Then, they discussed producer behavior on the Norwegian shelf, concluding that the company would behave as an oligopolist, whereas other Norwegian producers would probably behave as price takers.

Based on these assumptions, they discussed the consequences for the company in relation to the Norwegian competitors and also to other producers. As mentioned above, this discussion was very detailed, revealing that particularly the two market analysts have a very thorough understanding of the behavior of producers such as the United Kingdom and The Netherlands. Then, the manager introduced the demand side, and there was a brief discussion related to a growth prognosis for gas demand in Europe, particularly the United Kingdom.

The conclusion of the first discussion was that the team believed that prices most probably would go down. They believed that volumes would be rather unchanged because the authorities would still be able to control production. They expressed doubts regarding the development of profit, believed that profits most probably would go down, but did not expect any dramatic changes.

An interesting aspect was that the conclusion appeared at the start of the session and was almost not changed during the discussion, but only moderated. For example, a statement such as: "The prices will go down," was modified to: "We do not believe that prices will go up, most probably they will go down." A few critical questions were posed in connection with the moderation of the conclusion, e.g., whether large producers might overflow the markets. The following discussion, however, was characterized by attempts to find arguments to support the conclusion that was generated at the start of the session.

5.3 Description of the Second Session (with VisualGas)

The session with VisualGas started with a discussion of the assumptions of the model and the expression of confidence that the model parameters gave a valid representation of the markets. All three participants had been involved in the parameterization of the model. Then, they generated a new base case, primarily changing the producer behavior for The Netherlands from price taker to oligopolist. They solved the scenario, and agreed that the results were reasonable after a discussion where they compared the model results with their perceptions of the current market situation.

As the next step they generated the scenario with the producer behavior they had concluded on in the first session, i.e., the company behaving as oligopolist and other Norwegian producers behaving as price takers. Then, they solved the scenario – and were rather surprised by the results. The profit was lower than expected, primarily due to much higher transportation costs than expected. This surprise started a search for explanation of the results by developing a series of scenarios: They changed the behavior of the company to price taker. They also removed a restriction on a pipeline capacity. Finally, they ran the scenario they considered the most plausible, namely

the company behaving as an oligopolist without the earlier imposed restriction on the pipeline capacity – and other Norwegian producers behaving as price takers.

In the evaluation of the outcomes of each scenario they studied various screen pages, comparing the effects with the outcomes of the previous scenario. They looked at the changes in trade patterns, utilization of pipeline capacity and shadow prices. The direction of the changes in model results was as expected, but they had problems understanding some of the effects, particularly the heavy increase in transportation costs.

The preliminary conclusion of the second session was to check the model assumptions regarding transportation costs. Then, they needed to run more scenarios before making up their minds regarding the consequences of the liberalization of the gas markets.

5.4 Comparison of the Sessions

Table 2 summarizes the above description of the two sessions.

The table shows that team members handled the task at a higher level of group development in the session with VisualGas (Schroder et al. 1967), i.e., team learning.

Table 2 Level of group development before and after use of VisualGas

Level of Group Development	Before	After
Biases		
Availability	Increased volumes, pressure on prices	All model variables taken into account
Anchoring	Yes	No
Search for confirmation	Yes	Yes, but did not get
Number of alternatives	1 (2)	5
Base case, current situation	Implicit	Modeled explicitly
Scen. 1	Norway 1 N-C/Norway 2 PT	Norway 1 N-C/Norway 2 PT
Scen. 2		Norway 1 PT/Norway 2 PT
Scen. 3		Norway 1 PT with removed restriction
Scen. 4		Norway 1 N-C with removed restriction
Comparing interpretations	Implicitly	Explicitly
Combining interpretations	No	Yes, search for explanations
Perception of uncertainty	Yes, volumes (supply)	Yes, profit, trade pattern, transp. costs
Level of evaluation	Ordinal	Ratio
Generation of conflict	Yes, but few critical questions	Yes, evidence against expectations
Handling of conflict	Argumentation to support conclusion	Attempts to understand model behavior
Closure (conclusion)	Yes	No

$N - C$ Nash–Cournot oligopoly, *PT* price taker

In the second session the tendency to use simplifying heuristics, such as availability and anchoring (Tversky and Kahneman 1974), was reduced. In the first session focus was on gas supply and prices. In the second there was more focus on profit as a result of considering prices, income and costs. Much attention was paid to variables such as trade pattern and transportation costs, i.e., variables that are related to the interaction of supply and demand. Without a computerized tool to perform the necessary calculations it is difficult to evaluate how changes in market assumptions affect profit. In the first session there was a tendency to anchor the effects of the liberalization to the current market situation. The team did not expect volumes, prices and profit to change dramatically. In the second session, however, VisualGas calculated and presented the effects of the same scenario compared to the outcomes of a base case that they had just considered to be a fair representation of their own beliefs. The differences between the calculated effects and the team expectations were dramatic, and VisualGas thus prevented an anchoring effect. Instead of searching for confirmation, they had a need to search for explanation of the surprise.

In the first session the members only generated one set of consequences. They did not state their assumptions of the current market situation explicitly, so comparisons were only implicit, and there were no alternative scenarios to combine. In the second session they had to start with the generation of a case representing their current market assumptions, because VisualGas requires a base case for comparison of scenarios. Since the equilibrium solution was rather surprising, they generated four additional scenarios in a search for an explanation as described above. The outcomes of each scenario were explicitly compared to the outcomes of the previous scenarios. The outcomes of the second and third scenarios were also combined to form a hypothesis that was only partly supported in the fourth scenario.

In the first session the subjects evaluated consequences of the liberalization mainly using general economic theory. Their evaluation of the *direction* of the changes in key variables such as profit, prices and volumes was as calculated by VisualGas, but their evaluations were at the ordinal level ("profit will be reduced"). In the second session, however, VisualGas made it possible for them to use a ratio level in the evaluation of the consequences ("profit will be reduced by x%"), and their perception of environmental uncertainty related to the market change was increased.

In the first session there was a tendency to minimize conflict and to search for arguments that supported the consequences stated at the beginning of the session. This finding indicates that the team fell into the confirmation bias trap (see, e.g., Bazerman 1998). Thus, the team was able to reach a conclusion after one and a half hours. In the second session the subjects' perceptions of the uncertainty of the outcomes increased, and they were not able to reach a conclusion regarding the consequences of the market event at the end of the 2 h scheduled for the session.

The second session started with the scenario that was concluded on in the first session, i.e., the scenario the team considered the most probable. However, the consequences computed by VisualGas were very different from the team members' expectations developed in the first session. The above descriptions show that the subjects did not use any arguments from the first session in their evaluations of the

consequences of the market event in the second session. It is, therefore, not plausible that the differences between the two sessions were caused by learning in the first session. Furthermore, the second session took place 1 h after the first session, and we had lunch with the subjects during the break. There were, therefore, no "outside forces" that could have affected the results.

6 Concluding Remarks

In this paper we have presented the results of a quasi experiment to explore the effects on team learning of using an equilibrium model, VisualGas. The findings reported above demonstrate that such models can support managers/analysts in developing a more varied interpretation of environmental changes, i.e., enhance learning:

- Taking a larger number of relevant aspects into consideration;
- Operating with a more varied set of assumptions and evaluating a larger number of consequences for the company;
- Evaluating consequences using quantitative (ratio level) expectations instead of ordinal, thereby increasing their sensitivity to how critical the market event is to goal attainment.

Equilibrium models can, however, only help the users calculate and present results based on the variable and parameter values they enter into the system. The users must have a thorough knowledge of the task environment, and they need other information sources to help them notice environmental changes and to generate plausible scenarios. They must also understand both the possibilities and the limitations of the modeling format, and they must compare model assumptions and outcomes with empirical data to check whether the model is a valid representation of their task environment. Therefore, the usefulness of equilibrium models depends heavily on the participation of experienced managers/analysts in the user teams.

Using the system, however, there was a tendency that the subjects limited their attention to the model variables. In the first session the subjects generated more ideas based on their detailed knowledge of individual producers – but they were not able to incorporate their ideas in their evaluations of the consequences of the event. Furthermore, even though the team members considered more scenarios in the second session, the additional scenarios represented systematic changes in assumptions to understand model results related to the scenario they considered the most probable. They did not involve, e.g., assumptions of changed behavior of other European producers as well. This lack of exploration of alternative scenarios may, however, be due to the fact that the subjects had problems making sense of the increase in the transportation costs within the time scheduled for the session.

This paper reports on the use of one particular equilibrium model in a specific research setting. More research is called for to investigate the interaction between equilibrium models and human beings to enhance learning. In our quasi experi-

ment one of the strengths of using the model was that it generated a surprise, stimulating a search for explanations (Weick 1995; Louis 1980) – that also raised questions regarding the validity of some of the model assumptions. Longitudinal studies should be performed to investigate the interplay between the users' detection of changes in the task environment and the changes in/updates of the task model, i.e., the interaction between the users' knowledge and the *representation* of part of this knowledge in external organizational memories. Furthermore, longitudinal studies should be undertaken to find out whether use of equilibrium models enhances the development of users' knowledge structures, increasing their ability also to detect and make sense of weak environmental signals – and thus also a more proactive and effective handling of complex events in ever-changing environments.

References

Anderson, J. R. (1995). *Learning and memory*. New York: Wiley

Argyris, C., & Schön, D. A. (1996). *Organizational learning II: Theory, method and practice*. Reading, MA: Addison-Wesley

Bazerman, M. (1998). *Judgment in managerial decision making* (4th ed.). New York: Wiley

Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management, 23*(3), 239–290

Drucker, P. F. (1974). *Management: Tasks, responsibilities, practices*. New York: Harper and Row

Easterby-Smith, M., Snell, R., & Gherardi, S. (1998). Organizational learning: Diverging communities of practice? *Management Learning, 29*(3), 259–272

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence in the illusion of validity. *Psychological Review, 85*, 395–416

Fuglseth, A. M. (1989). *Decision support: Methods for diagnosis of managers' information and situation perceptions (in Norwegian)*. Unpublished PhD dissertation, Norwegian School of Economics and Business Administration, Bergen, Norway

Gherardi, S., Nicolini, D., & Odella, F. (1998). Toward a social understanding of how people learn in organizations: The notion of situated curriculum. *Management Learning, 29*(3), 273–297

Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. *Organization Science, 2*(1), 88–115

Katzenbach, J. R. (1998). *Teams at the top: Unleashing the potential of both teams and individual leaders*. Boston, MA: Harvard Business School Press

Louis, M. R. (1980). Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25*(2), 226–251

Mathiesen, L. (1985). Computational experience in solving equilibrium models by a sequence of linear complementarity problems. *Operations Research, 33*(6), 1225–1250

Mathiesen, L. (1987). *GAS: En interaktiv modell og et interaktivt system for analyser av det vesteuropeiske marked for naturgass (GAS: An interactive model and an interactive system for analyes of the Western European natural gas market)*. SAF-Report No.3, Centre for Applied Research, Norwegian School of Economics and Business Administration, Bergen, Norway

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14–37

Pawlowsky, P. (2001). The treatment of organizational learning in management science. In M. Dierkes, A. Berthoin Antal, J. Child & I. Nonaka (Eds.), *Handbook of organizational learning and knowledge* (pp. 61–88). Oxford: Oxford University Press

Schroder, H. M., Driver, M. J., & Streufert, S. (1967). *Human information processing – individuals and groups functioning in complex social situations*. New York: Holt, Rinehart and Winston

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Bantam Doubleday

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science, 2*(1), 125–134

Tversky A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, New Series, 185*(4157), 1124–1131

Walsh, J. P., & Ungson, G. R. (1991). Organizational memory. *The Academy of Management Review, 16*(1), 57–91

Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. New York: Cambridge University Press

Refinery Planning and Scheduling: An Overview

Jens Bengtsson and Sigrid-Lise Nonås

Abstract In this chapter, we give an overview of recent literature on the planning and scheduling of refinery activities. Planning of refinery activities ranges from determining which crude oil types to acquire to which products that should be produced and sold in the market. The scheduling ranges from scheduling of crude oil unloading and blending to blending of components to finished products. This overview treats three different categories of activities: planning and scheduling of crude oil unloading and blending, production planning and process scheduling, and product blending and recipe optimization. The focus will be on identifications of problems, the models outlined for the specified problems, and the computational difficulties introduced by the models. A final section discusses an agenda for future research.

1 Introduction

Building a modern refinery is a huge investment that puts its owner in a position where high fixed cost must be covered for a lengthy future. Because of this investment and fixed cost, efficient use of refinery resources is important both for short-term and long-term profitability. In addition, refineries incorporate complex equipment and produce complicated chemical reactions, posing difficult challenges for determining the best use of the refinery capacity.

S.-L. Nonås (✉)
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: sigrid-lise.nonas@nhh.no

J. Bengtsson
Department of Accounting, Auditing and Law, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: jens.bengtsson@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_8,
© Springer-Verlag Berlin Heidelberg 2010

Apart from the complexity of the refinement and the challenges of determining efficient processes within the refinery itself, other factors play key roles in the search for profits within the industry. The markets for crude oil and petroleum products have developed over the decades and have shown to be sensitive both to political issues and disruptions and variations in demand and supply. The market for oil related products is well-developed and this also includes the freight market. The latter is important since crude oils and products are dependent on transport to their intermediate and final destinations. Accordingly, profitability will also be affected by transportation costs.

The supply chain of an integrated oil company stretches from the production and purchasing of crude oil to customers buying petrochemical products or fuel for heating or transport. There are many decisions that must be made along the supply chain such as what crude oil mix to buy and sell, what components and products should be produced, and whether they should be kept for internal use, stored, or sold to external players.

To address the challenges in the oil and gas supply chain quantitative models and mathematical programming techniques have been developed for several decades and their use has significantly increased the ability to plan and control refinery activities and to increase profits.

From a refinery management perspective there is a difference between planning and scheduling. At the planning stage, the time horizon is typically several weeks or months, and the decisions typically concern purchase of crude oils and the production and sales of products. Since the markets associated with refinery operations are volatile, the use of correct and updated information is important because this will strongly affect the capability to identify market opportunities. The identification of market opportunities is crucial for increased profitability. The capability of identifying market opportunities will be dependent on a company's ability (given current condition: prices, production decisions, available crude, etc.) to determine its decision to buy, refine, and sell its products. To make such decisions, the company must consider already booked and planned production, together with future prices.

Due to the complexity involved in different refinery operations throughout the supply chain, the refinery scheduling problem is often separated into three different sub-problems, see Fig. 1 below. The first sub-problem involves the crude oil unloading from vessels or pipelines, its transfer to storage tanks and the charging schedule for each crude oil mixture to the distillation units. The second sub-problem consists of production unit scheduling, which includes both fractionation and reaction processes. The third sub-problem is related to the blending, storage, and lifting of final products.

Historically, refiners have built organizations based on the processes associated with planning and scheduling. To drive operational efficiency, major refining companies are now putting increased focus on managing supply chain activities as an integrated process, closely connecting refinery planning and scheduling to improve communication and total plant operation.

In this chapter, we will give an overview of the latest literature on refinery planning and scheduling and also make some suggestions for future research. A previous

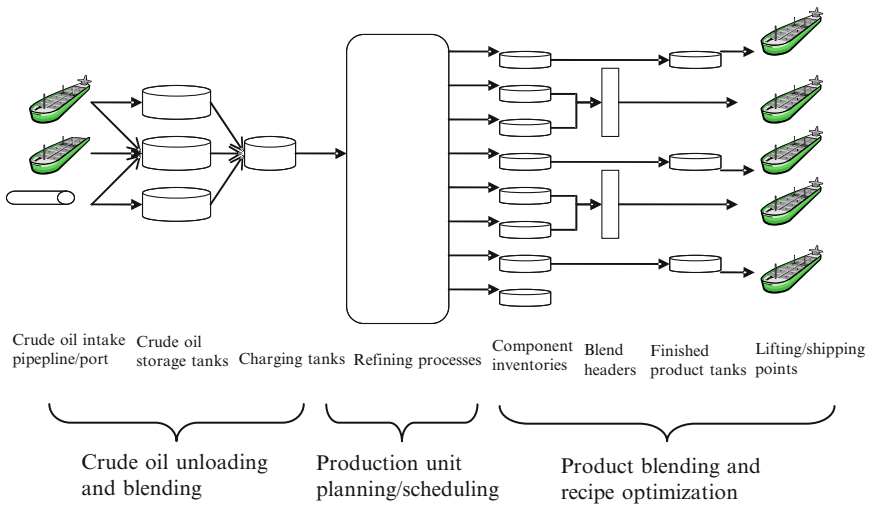


Fig. 1 Three sub-systems of an oil refinery

survey of literature on production planning and scheduling models for refinery operations can be found in Pinto et al. (2000). The focus of the chapter will be on the identifications of problems, the type of models outlined to solve the identified problems, and the computational difficulty introduced by the models (it will not discuss how the computational difficulties are met). The overview is organized in three different parts, which correspond to the sub-problems mentioned earlier. Literature that focuses on modeling the whole refinery supply chain is also discussed. This can be found under the production planning section.

2 Crude Oil Selection and Crude Oil Scheduling

The planning and scheduling of crude oil operations in a refinery is a critical task that can save the refinery millions of dollars per year (Kelly and Mann 2003a, b). Crude oils vary significantly in compositions, product yields, properties, and prices, and their acquisition accounts for a large portion of the refineries' cost. A key issue for a refinery is, therefore, to identify and process optimal crude blends that maximize profit margins (Li et al. 2007).

Typically, an oil refinery receives its crude oil through a pipeline that is linked to a docking station, where oil tankers unload. The unloaded crude oil is stored at the refinery in crude oil tanks. The crude oils are stored in these tanks, at least for a minimum amount of time (to allow the separation of the brine), before the mix of crude oils is transferred to charging tanks or directly blended and processed in the distillation tower.

In general, two types of ships supply crude to a refinery: very large crude carriers (VLCCs) that may carry multiple parcels of different crudes and small vessels

carrying single crude. Due to its size the VLCCs often dock at a single buoy mooring (SBM) station offshore whereas the small vessels berth at the jetties. The pipeline connecting the SBM station with the crude tanks normally has a substantial holdup (Li et al. 2007), while the holdup in the jetty pipeline is not that critical.

The shipping schedule for the crude oil tankers is determined by the procurement and logistic department in conjunction with the crude oil purchase. Due to lengthy waterborne transit times, this schedule is done a long time before the crude oil tankers arrive at the refinery. In the case that the crude oil carrier transport different crude oil parcels, the unloading sequence is also predetermined due to logistic considerations. Before a VLCC can unload, it must first eject the crude that resides in the SBM pipeline. This crude can, as suggested in Reddy et al. (2004b), be modeled as an extra crude oil parcel from the carrier. A general assumption is that the holdup of the jetty pipeline is negligible. In the literature, it is typically assumed that the holdup of the SBM pipeline is negligible.

The crude oil scheduling is based on the current information on arrival of crude oil vessels or carriers, crude composition in different tanks, and the optimal crude feed to the crude distillation units (CDUs) from the production plan.

2.1 Selection of Crude Oils

The main goal of the crude selection is to find a feasible crude blend that maximizes the profit of the planned production in the time horizon considered, taking into account the current storage of crude oil at the refinery and the crude tankers scheduled to arrive at the refinery. The "wrong" crude mix can cost a refiner both in excess feedstock expense and lost product revenue. To find the right crude mix the scheduler has to take into account both processing and economic considerations. After the selection of crude oils, the crude procurement and logistics departments have to secure the crudes and schedule them for delivery.

The monthly planning model, updated with more advanced crude blending simulations, is often used as a decision tool in order to determine the optimal crude oil mix. Each refinery has built up its own history of how different crudes have performed for their refinery in the past. This information is then used in the crude blending simulation to achieve an optimal crude slate. Both Stommel and Snell (2007) and Ishizuka et al. (2007) present good discussions and describe general rules for how leading companies handle the selection and the scheduling of crude oil.

In the literature, crude oil selection has been considered as part of the production planning problem. The literature, however, neglects one important component. The complex nonlinear blending relation of different crude oil mix is omitted in the planning problem due to the increase in problem complexity. In the industry, these relations are usually solved using in-house or more commercial simulation tools. The general review of literature related to production planning is presented in a later section.

2.2 Scheduling of Crude Oils

The objective of crude oil scheduling is to minimize the operational cost while meeting the target crude oil load to the CDU. For the scheduling one assumes a fixed arrival schedule for vessels, knowledge of the quantity and quality of crude oil at the vessels and the crude oil mix at the crude tanks, the minimum settling time for the brine at the tanks, minimum and maximum level of crude in the tanks, minimum and maximum flow rate in the pipeline, and the target feeding-rate (quality and quantity) of the crude oil blend for the CDU. It is also common to assume that a tank cannot receive and feed crude at the same time.

Given these facts, the objective of the crude oil scheduling is then to determine the unloading schedule for each vessel (this includes the timing, rates, and which tank to transfer the oil parcels to); the transfer schedule of crude oil mix between storage and charging tanks (if both are present in the refinery); the inventory levels and crude oil concentration for all storage and charging tanks; and finally the charging schedule for each CDU (how much should be transferred to each CDU from each charging tank).

To model the crude oil scheduling we need a large number of binary allocation variables to consider the discrete scheduling decisions, such as selecting a tank to receive a crude oil mix and nonlinear constraints to calculate the crude oil composition for the storage and charging tanks. So far, the models proposed for crude oil scheduling have not considered nonlinear crude properties. The nonlinear property constraints are approached by linear constraints that consider key crude component concentrations or blending indexes that are linear on a volumetric or weighted basis. Even if we assume linear crude properties, the crude oil scheduling results in a complex mixed integer nonlinear (MINLP) model. The nonlinear terms are bilinear and originate from the mass balance and crude mix composition constraints for the storage and charging tanks and the feed from these tanks. A nonlinear term $f(x, y, z)$ is said to be bi-linear if it is linear with respect to each of its variables, i.e., $f(x, y, z) = x^*y + y^*z + z^*x$.

Several authors have discussed different models and methods to solve this crude oil blending and scheduling problem, taking into account different degrees of refinery complexity. The scheduling problem can be modeled using either a continuous time approach or a discrete time approach. The discrete time formulation tends to rely on an excessive number of time periods to achieve the required accuracy, while continuous time formulation results in complex models that require several assumptions or specialized algorithmic solution techniques. Recent trends in scheduling models for chemical processes have however moved toward continuous time formulations to avoid the high number of integer variables found in discrete time models. The continuous time modeling is particularly suited to crude oil scheduling since refinery activities can range from some minutes to several hours (Joly et al. 2002). A review that compares discrete and continuous time approaches of scheduling for chemical processes is provided by Floudas and Lin (2004).

One of the first models presented for the crude oil scheduling is a discrete mixed-integer linear programming (MILP) presented by Lee et al. (1996). The authors

consider one docking station, a set of storage tanks, a set of charging tanks and a set of CDUs The objective is to find a schedule that meets the predetermined crude slate for the CDU, while minimizing total operating cost (unloading cost, cost for demurrage, tank inventory cost, and changeover cost for the CDUs). In addition, the crude mix in the charging tanks should be within predefined quality measures with regard to key component concentration. Their linear reformulation of the bi-linear mass balance constraints is, however, not rigorous enough to ensure that the crude mix composition for the storage and charging tanks is the same as the composition of the flow from the tanks. This inconsistency is denoted as "composition discrepancy" by Li et al. (2002) and Reddy et al. (2004a). In general, composition discrepancy may occur when individual component flows are used in a linear reformulation for the bi-linear mass balance term for storage and charging tanks where mass accumulates. Li et al. (2002) proposed a discrete mixed integer nonlinear programming (MINLP) model that extended the model in Lee et al. (1996) by reducing the number of discrete decision variables (they replace two bi-index binary variables with one tri-index binary variable) and by including new features as multiple jetties and allowing the possibility for two tanks to feed a crude distillation unit. The solution approach outlined for the MINLP model may, however fail to find a feasible solution even if one exists (Reddy et al. 2004a; Kelly and Mann 2003a, b).

Jia and Ierapetritou (2003, 2004) outline a continuous time MILP model for the crude oil scheduling, using only linear terms for the crude mix operations as in Lee et al. (1996). They assume no cost for crude or tanks changes, one tank feeding one CDU at a time (and vice versa). The objective is to minimize demurrage, unloading and inventory cost. As for Lee et al. (1996), the MILP model may suffer from composition discrepancies. The crude composition in the storage and charging tank may not match its feed to the charging tanks and to the CDU.

Moro and Pinto (2004) propose a continuous time MINLP formulation for the crude oil scheduling. The model considers one pipeline to unload the crude oil, settling time in the crude tanks to separate brine from the oil, and at most, two crude oil tanks feeding the CDU. They propose to measure the quality of the CDU stream by limiting the concentration of the critical components in the feed to the CDU. Their objective is to maximize the CDU feed rate while minimizing the crude tank operating costs. Moro and Pinto (2004) also proposed a MILP approach of the MINLP model where the bilinear crude mixing term is linerarized by discretizing the amounts or types of crude oils present in the storage tanks. The MILP approach suffers, however, from an increasing number of binary variables as the number of discretization interval increases. Reddy et al. (2004a) developed, in parallel to Moro and Pinto (2004), a continuous time MINLP formulation considering multi-parcel vessels loading at one SBM, pipelines that transfer the crude parcels from the SBM to the crude storage, and charging tanks that again feed the CDUs. Multiple storage, tanks can feed one CDU (and vice versa). They also discuss how SBM parcels can be created to take into account that before a vessel can unload, the crude in the SBM pipeline (the SBM parcel) has to be transferred to a storage tank. The objective of the scheduling model is to maximize gross profit (profit of products-cost of crude oils) while minimizing the operating cost (CDU changeover cost, demurrage, penalty for running under the crude safety stock).

In addition to the features considered in the continuous time model the discrete-time model outlined by Reddy et al. (2004b) accounts for multiple vessels unloading to a set of jetties and transfer of crude oil between the storage tanks. They also present a novel approach for dealing with parcel unloading, which uses fewer binary variables than earlier work (Li et al. 2002; Lee et al. 1996). In addition, they allow more than one unloading allocation in any time period, thus utilizing the entire time period to the maximum extent. The objective for their scheduling model is to maximize the crude margin (total value of cuts from the crude distillation units minus the cost of purchasing, transporting, and processing the crude) minus the operating costs (changeover, demurrage, penalty for running under the crude safety stock). The unloading cost and inventory cost is not considered since the amount of crude is fixed for the scheduling horizon. The demand for crude oil for each CDU has, however, to be satisfied. Their solution approach improves the solution approach proposed by Li et al. (2002) but may still fail to find a feasible solution even if one exists. Li et al. (2007) improved the MINLP formulation of Reddy et al. (2004a) in two ways. First, constraints that disallow uncontrolled changes in the CDU feed rates are inserted. Second, linear blending indexes, weight- and volume based, are used to better approximate the nonlinear crude oil properties.

From this literature we see that the crude scheduling problem is approached with both MILP and MINLP models and solved either with standard MILP and MINLP solvers or with tailor-made solution approaches. A global optimization algorithm for the crude scheduling problem would be preferable, but considering the large sizes of practical problems and the need for quick solutions, that will require considerable effort and is a great challenge for future research. The main difficulty is to deal with the large number of integer variables in the model and the complexity of nonlinear blending and crude oil mixing operations.

Note that disruptions such as crude arrival delay could make any given schedule infeasible and necessitate rescheduling of operations. Adhitya et al. (2007) discuss how to reschedule in order to make only minimal changes to the scheduled operations rather than undergo a total reschedule.

3 Production Planning and Scheduling

The refinery is built up of different processing units that transform a variety of input streams into several output streams. The flow rate and product specification, e.g., octane number and sulphur content, of each output stream is determined by the flow rate and product specification of the unit feed and the operating mode of the processing unit. Nonlinearity arises from mixing the feed and, in the yield, from the processing. The change of operating mode of a processing unit results in a period with reduced capacity and disturbance in the yield (quality and quantity) of the processing. Reduced capacity and disturbance in the yield also occur for a start up after a period with maintenance. To correctly model the disturbance caused by a change in operating mode is difficult, usually the issue is relaxed by incurring a setup cost for each change in mode.

For a general refinery, the planning specifies which crude or intermediate products to purchase and which products to produce. The planning decisions are taken based on forecast of future demand, and usually planning takes into account a two or three months time horizon. The production plan is used in a rolling horizon setting to take into account updated information regarding refinery and markets. The decisions related to scheduling of refinery activities are generally performed on the basis of shorter time horizons, i.e., days or weeks, to determine the specific timings of the operations.

3.1 Production Planning

To support production planning decision making, refineries generally use commercial packages based on linear one-periodic models that rely on constant yields (PIMS from Aspen Tech and RPMS from Honeywell, Process Solutions). This has motivated researchers to outline models that give a more accurate representation of the refinery processes or activities. In this section, literature that takes into accounts multiple periods and different degrees of nonlinearity in the mixing and processing operations are presented. Papers that consider a supply chain perspective and uncertainty in market data are also presented.

One of the first contributions to consider nonlinearity in the production planning is that of Moro et al. (1998). Moro proposes a framework where every unit in the refinery is represented as an entity and the complete refinery topology is defined by connecting the unit streams. For the processing units nonlinearity can be considered in the blending relations and in the process equations. A general MINLP model is discussed for a diesel production planning problem, but this is only partly outlined in the paper. Detailed blending and processing formulations are presented only for the hydro-treating unit. They report that the refinery plan obtained from the MINLP model improved the performance of the case company significantly compared to the current operating decision that was based on experience and manual calculations. The same planning model is discussed in Pinto and Moro (2000), here with results from a new case study.

Neiro and Pinto (2005) formulate a MINLP model that extends the planning model discussed in Moro et al. (1998) to account for multiple time periods and uncertainty in market data. Uncertainty is considered in the product demand, the product price and the cost of crude oil. The uncertainty is expressed in scenarios, and the objective function includes weighted values of each scenario based on the probability for each scenario to occur. For each time period, the main decisions are which crude oil to select, how to operate the processing units, and how much of the final products to hold in inventory. They show an exponential increase in solution time with the number of time periods as well as with the number of scenarios. In the work listed above, only sub-systems of the gas and oil supply chain have been considered in a reasonable level of detail. Neiro and Pinto (2004) propose a framework for modeling the whole petroleum supply chain, including crude oil suppliers, distribution centers, and several complex conversion refineries interconnected by

intermediate and end product streams. This study outlines a large scale one periodic MINLP planning model for the system addressing crude oil purchasing, production units processing, inventory management, logistics, and end product sales decisions. Neiro and Pinto (2004) consider nonlinear blending for the different processing units and storage tanks, and nonlinear operating conditions in accordance to the yield from the processing units. They consider a supply chain with four refineries connected with pipelines and storage tanks, each with different capacity and topology.

Refinery planning and the refinery scheduling are generally performed sequentially, mainly due to the complexity of the refinery sub-problems. When determining the refinery production plan, many of the scheduling constraints are uncertain. To obtain a feasible schedule which utilizes resources in, or close to, an optimal fashion, it is important that the company determine a plan which makes this possible. If the planning and scheduling is done sequentially, there is no guarantee that the production plan can give an operable schedule. Kuo and Chang (2008) have addressed this issue and present a MILP planning model that addresses stream allocations and processing run modes for several scheduling intervals. By considering the whole refinery supply chain and splitting the planning period into several sub-intervals, Kuo and Chang, are better able to match the planning and scheduling decisions and improve the performance of the supply chain scheduling activities.

Environmental regulations and the risks of climate change pressure refineries to minimize their greenhouse gas emissions. Refineries also face more stringent product specifications on oil products which typically increase their energy consumption and CO_2 emissions. Szklo and Schaeffer (2007) address this problem, also with a specific focus on the Brazilian refining sector. Holmgren and Sternhufvud (2008) discuss different possibilities for reduction of CO_2 emissions for petroleum refineries in Sweden. More analytical approaches to this problem have also been addressed. Babusiaux (2003), Nejad (2007), Pierru (2007), and Babusiaux and Pierru (2007) have proposed different methods for allocating the CO_2 emissions among the different refinery products produced.

Elkamel et al. (2008) propose a MILP for the production planning of refinery processes. They consider how to find suitable CO_2 mitigation options for the processing units that meet both a CO_2 emission target and the final product demand while maximizing profit.

Zhang and Hua (2007) propose a MILP model for a multi-period planning model that considers the integration of the processing system and the utility system for the refinery industry. The objective here is to determine an optimal material and energy flow in order to maximize the overall profit.

Uncertainty is present in different forms in the different sub-system in the refinery. Some of the latest work that considers uncertainty in the planning and scheduling of refinery activities is by Neiro and Pinto (2005, 2006), Pongsakdi et al. (2006), and Zimberg and Testuri (2006). Pongsakdi et al. (2006) address the planning of crude oil purchasing and its processing based on the network structure proposed by Pinto et al. (2000). Uncertainty is considered both in product prices and product demands. The stochastic problem is modeled as a linear two stage stochastic model with recourse and is solved using a special implementation of the average

sampling algorithm introduced by Aseeri and Bagajewicz (2004). Test results show that in comparison to the stochastic solution the deterministic solution has a lower expected gross refinery margin and a larger risk. The profit is maximized taking into account product revenues, crude oil costs, inventory costs, and cost of unsatisfied demand.

Zimberg and Testuri (2006) consider the crude oil procurement and processing for a case company that has a specific focus on the bunker fuel oil production. Generally, the gas oil, diesel, and fuel oil products are more dependent on the right crude oil mix than the gasoline products. A simplified two stage stochastic process is considered. The first stage decision, purchasing of crude oil, is taken two months before the second stage decisions, the processing of the crude oil and blending of intermediate products. A case study that compares the deterministic (mean values) and stochastic solution of the problem is presented. The risk is not considered, only the expected refinery margin. Neiro and Pinto (2005, 2006), as previously discussed, take into consideration uncertainty in product demand, product prices and crude oil cost. The uncertainty is modeled in discrete scenarios and weighted according to the probability of occurrence.

Pitty et al. (2008) and Koo et al. (2008) develop a decision support for an integrated supply chain, which in this case means that it handles activities such as crude oil supply and transportation in addition to intra-refinery activities. The decision support is based on a simulation-optimization framework and takes stochastic variations in transportation times, yields and prices into account. Pitty et al. present the complete dynamic model whereas Koo et al. use the decision support to optimize the design and operation in three different supply chain related cases.

Current research has to a large degree been focused on modeling and analyzing different types of refinery planning problems and has used commercial solvers as GAMS (OSL, DICOPT, CONOPT) with the different solution approaches they offer to solve the outlined problems. Nonlinearity is considered in some degree and for some problems multiple periods is proposed. Also, a supply chain view has been considered. Some recent papers consider decomposition strategies to solve the large complex nonlinear planning problem. Future research should consider methods for solving the complex MINLP problems more efficiently and focus on more advanced methods that consider uncertainty in market data.

3.2 Production Scheduling

In the literature, MILP models are generally outlined for production scheduling problems. The models focus on part of the refinery activities and incorporate different degree of details regarding the blending and processing in the processing units. The models are generally solved using standard commercial solvers as GAMS (CPLEX, OSL, DICOPT, CONOPT). Pinto et al. (2000) propose a discrete MILP model for the production and distribution scheduling of diesel. The scheduling

system considered a set of crude distillation units that produce specified intermediate (or end) products for storage in a set of intermediate tanks before they are mixed and sent through pipelines to supply the consumer market where there is a known demand. The proposed MILP model considers linear blending relations. Results are reported for a case company considering market demand for three different types of diesel oil. A scheduling horizon of one day considering one hour time intervals is addressed.

Joly et al. (2002) outline both a discrete MINLP model and a discrete MILP model for the scheduling problem that considers decisions related to mixing, storage and distribution. The system configuration includes 1 deasphalting unit, 1 cracking unit, 2 diluents storage tanks, 15 storage tanks for final products, and 2 pipelines. The nonlinear terms in the MINLP model refer to the calculation of viscosity in the oil mix and are linearized in the MILP model. The models produced similar results in terms of solution quality, while the solution time for the MILP model was a bit longer. Moro and Pinto (2004) and Joly et al. (2002) also discuss a scheduling problem that addresses how to make use of the given raw materials, processing units, and storage in order to meet the LPG product deliveries. They did not investigate formulations related to product quality.

Goethe-Lundgren et al. (2002) outline a discrete MILP model for a refinery production scheduling problem considering one distillation unit and two hydrotreatment units that can each operate in 5–10 modes. The model considers how to run the processing units in an optimal manner to minimize production and inventory costs. The model can also be used to evaluate the feasibility and the production cost for given product and crude oil shipping plans. To make the base schedule robust, Goethe-Lundgren et al. (2002) implemented new constraints in the model that assure that enough end products are available if a tanker should arrive one day too early, and to assure enough storage capacity if a tanker is delayed one time period. They also report how the flexibility decreased (the cost increased) when a more robust schedule was offered.

Jia and Ierapetritou (2004) propose continuous time MILP formulations for specific crude oil, production and product scheduling problems. A lube-oil refinery is considered for the processing units parts. The study proposes a continuous time MILP formulation for the scheduling problem that takes into account material balance in tanks and production units; capacity, allocation, and sequence constraints; and demand constraints.

Due to the complexity of the problem, the current work proposed for the production scheduling relaxes most of the nonlinear relations, and only simple refinery systems or sub-systems of the refinery topology are considered. Future work in this area should focus on formulating models and finding corresponding solution approaches that enable companies to solve nonlinear production scheduling models of real-world sizes in a reasonable time. In addition, new work that studies how the daily scheduling decisions might best be incorporated in the production planning and how uncertainty in the market data could be modeled in the scheduling problem would also benefit the gas and oil industry.

4 Product Blending and Recipe Optimization

The product blending problem concerns the best way to mix different components and additives in order to meet quality and demand requirements of the final products. Several final products can be produced out of a number of combinations of components, but some properties of the final product do not show linear relationships, e.g., pour point of diesel. These relationships put requirements on the optimization model so that the nonlinearity must be handled in some way.

The product blending problem is usually split into a long-term problem and a short-term problem. In the case of long-term situation, the problem is basically to determine optimal recipes that maximize profit given quality specifications and quantity requirements. In the short-term situation, detailed operational and temporal constraints come into play and the basic issue becomes that of scheduling.

Glismann and Gruhn (2001) claim that short-term scheduling of blending processes is more complicated than scheduling of most other processes because of the option of blending a product in many different ways. In the scheduling of blending, consideration of recipe optimization and short-term scheduling within an integrated approach becomes necessary. To address this Glismann and Gruhn develop an integrated approach to coordinate nonlinear recipe optimization and short-term scheduling of blending processes. They propose a two-level optimization approach that sets up a large scale NLP model to determine product quantities and recipes. Given the result from the NLP model, a MILP model based on a resource-task-network is used for the scheduling problem to optimize resource and temporal aspects. Both models are discrete time models. Whereas the NLP model maximizes profit, the MILP model minimizes deviations from given tank volume goals. Glismann and Gruhn also present alternative strategies to handle situations where a given goal cannot be met. They advocate integrating the planning and scheduling by using an iterative approach so that if the goal at the scheduling level cannot be met (due to bottlenecks), then the new information will be brought back to the planning level and the modified NLP problem solved again. After this, the MILP problem would be reconsidered until a feasible or satisfying solution is found.

Jia and Ierapetritou (2004) consider scheduling of gasoline blending and distribution as one of three sub-problems. The other two sub-problems consider crude oil and processing units. They assume that the recipe of each product is fixed, in order to keep the model linear. The problem is modeled as a MILP-problem in continuous time and based on the state-task-network (STN) representation introduced by Kondili et al. (1993). In Jia and Ierapetriou the objective function is formulated such that certain flows are as close as possible to a goal flow, but they also mention that other objective functions can be used.

Mendez et al. (2006) point out that there are a number of mathematical programming approaches available to the short-term blending and scheduling problem. But in order to reduce problem difficulty, most of them rely on special assumptions that generally make the solution inefficient or unrealistic for real world cases.

Some of the common simplifying assumptions are (a) fixed recipes are predefined, (b) components and production flow rates are known and constant, and (c) all product properties are assumed to be linear.

Mendez et al. develop a novel iterative MILP formulation for the simultaneous optimization of blending and scheduling and formulate the problem both in discrete and continuous time. The components flow from the processing unit is stored in component tanks before the components are blended in blend headers and sent to product tanks. The resulting nonlinear MINLP blending and shipment problem is modeled as a successive MILP problem. The objective function maximizes profit and is based on the assumption that the cost of components can be observed or determined.

Mendez et al. also highlight the fact that the multi-period product blending and product shipping problem is a complex and highly constrained problem where feasible solutions may be difficult to find. To increase the speed of the solution procedure, preferred product recipes could be included in the problem to help find a feasible solution more quickly. To avoid infeasible solutions, Mendez et al. propose to include penalty for deviation from preferred product recipe and penalties for deviations from specified product qualities. They also propose to allow purchase of components at a higher cost from a third-party to relax minimum inventory constraints.

Future work might focus on how to determine the component value and the preferred product receipt, in order to optimize the combined performance of the short-term blending and product shipments and how to coordinate the scheduling decisions with long-term planning decisions.

In the literature, the values of the components are commonly presented as a known value, as in Mendez et al. (2006). Often, however, the refinery's value of a certain component is unknown due one of two reasons; the component's value cannot be found from an external market or the value is not appropriate since lead times makes this option infeasible. A variety of methods, based on marginal values of components and properties and product values, can be used to determine the value of the components, and special attention should be made to the value that is used because different component values can give different optimal blends.

We know that the short-term blending decision is affected by two facts: (a) flexibility in the short-term is restricted, making it sometimes hard to stick to the recipe, which is considered optimal in the longer term, typically given by the planning model and (b) the relative value of blending components at the blending point in time might be different from the value determined in the long-term optimization. Thus, in the short-term, another recipe may be more profitable than the long-term optimal recipe, and the deviations from the long-term recipe may indicate that other blending recipes should be used. In the ideal world, it would be possible to observe the values of components, and in order to more closely approximate the values of these components, there must be integration between short-term and long-term decision.

5 Discussion and Further Research

This chapter has analyzed papers that consider planning and scheduling problems in refineries. The papers consider planning and scheduling problems mainly for refinery sub-systems and for refinery supply chains, in different forms and with different degrees of detail. It has been common to use commercial MILP and MINLP solvers to address the refinery model proposed. In addition, specialized algorithms have been proposed to solve specific industry sized problems in a reasonable time. Unfortunately, no general solution technique has so far been outlined that can solve real world problems in a satisfactory manner.

Due to the complexity of the refinery planning and scheduling problem, the works proposed to this date relax most of the nonlinear relations. Future work should focus on formulating correct NLP that take into account all aspects of the refinery sub-systems and on developing solution techniques that enable companies to solve nonlinear scheduling models of real-world sizes in a reasonable time.

Better coordination of refinery activities can increase throughput, reduce quality give away and demurrage, and improve the refinery's ability to evaluate special opportunities that may arise. Future research should consider how to coordinate the scheduling decisions with the long-term planning decisions.

Beyond developing solution techniques and more advanced models, there is currently an increased focus on environmental impact from activities associated with refining of crude oil. Tougher environmental regulations on oil products set by authorities and an increased focus on reduction of CO_2 emissions put new constraints and requirements on decision making and adds more complexity to an already complex situation. Given contemporary consensus about the environment, more research focusing on environmental impact is needed.

References

Adhitya, A., Srinivasan, R., & Karimi, I. A. (2007). Heuristic rescheduling of crude oil operations to manage abnormal supply chain events. *AIChE Journal, 53*, 397–422.

Aseeri, A., & Bagajewicz, M. J. (2004). New measures and procedures to manage financial risk with applications to the planning of gas commercialization in Asia. *Computers and Chemical Engineering, 28*, 2791–2821.

Babusiaux, D. (2003). Allocation of the CO_2 and pollutant emissions of a refinery to petroleum finished products. *Oil & Gas Science and Technology-Revue del Institut Francais du Petrole, 58*, 685–692.

Babusiaux, D., & Pierru, A. (2007). Modeling and allocation of CO_2 emissions in a multiproduct industry: the case of oil refining. *Applied Energy, 84*, 828–841.

Elkamel, A., Ba-Shammakh, M., Douglas, P., & Croiset, E. (2008). An optimization approach for integrating planning and CO_2 emission reduction in the petroleum refining industry. *Industrial & Engineering Chemistry Research, 47*, 760–776.

Floudas, C. A., & Lin, X. X. (2004). Continuous-time versus discrete-time approaches for scheduling of chemical processes: a review. *Computers and Chemical Engineering, 28*, 2109–2129.

Glismann, K., & Gruhn, G. (2001). Short-term scheduling and recipe optimization of blending processes. *Computers and Chemical Engineering, 25*, 627–634.

Goethe-Lundgren, M., Lundgren, J. T., & Persson, J. A. (2002). An optimization model for refinery production scheduling. *International Journal Production Economics, 78*, 255–270.

Holmgren, K., & Sternhufvud, C. (2008). CO_2-emission reduction costs for petroleum refineries in Sweden. *Journal of Cleaner Production, 16*, 385–394.

Ishizuka, I. T., Ueki, O., & Okamo, U. (2007). Improve refiner margins with integrated planning and scheduling systems. *Hydrocarbon Processing, 86*, 75–79.

Joly, M., Moro, L. F. L., & Pinto, J. M. (2002). Planning and scheduling for petroleum refineries using mathematical programming. *Brazilian Journal of Chemical Engineering, 19*, 207–228.

Jia, Z., & Ierapetritou, M. (2003). Mixed-integer linear programming model for gasoline blending and distribution scheduling. *Industrial & Engineering Chemistry Research, 42*, 825–835.

Jia, Z., & Ierapetritou, M. (2004). Efficient short-term scheduling of refinery operations based on continuous time formulation. *Computers and Chemical Engineering, 28*, 1001–1019.

Kelly, J. D., & Mann, J. L. (2003a). Crude oil blend scheduling optimization: an application with multi-million dollar benefits - Part 1. *Hydrocarbon Processing, 82*, 47–53.

Kelly, J. D., & Mann, J. L. (2003b). Crude oil blend scheduling optimization: an application with multi-million dollar benefits - Part 2. *Hydrocarbon Processing, 82*, 72–79.

Kondili, E., Pantelides, C. C., & Sargent, R. W. H. (1993). A general algorithm for scheduling batch operations. *Computers and Chemical Engineering, 17*, 211–227.

Koo, L. K., Adhitya, A., Srinivasan, R., & Karimi, I. A. (2008). Decision support for integrated refinery supply chains – part 2. Design and operation. *Computers and Chemical Engineering, 32*, 2787–2800.

Kuo, T. -H., & Chang, C. -T. (2008). Application of a mathematic programming model for integrated planning and scheduling of petroleum supply networks. *Industrial & Engineering Chemistry Research, 47*, 1935–1954.

Lee, H., Pinto, J. M., Grossmann, I. E., & Park, S. (1996). Mixed-integer linear programming model for refinery short-term scheduling of crude oil unloading with inventory management. *Industrial & Engineering Chemistry Research, 35*, 1630–1641.

Li, W. K., Hui, C. W., Hua, B., & Tong, Z. (2002). Scheduling crude oil unloading, storage, and processing. *Industrial & Engineering Chemistry Research, 41*, 6723–6734.

Li, J., Li, W., Karimi, I. A., & Srinivasan, R. (2007). Improving the robustness and efficiency of crude scheduling algorithms. *AIChE Journal, 53*, 2659–2680.

Mendez, C. A., Grossmann, I. E., Harjunkoski, I., & Kaboré, P. (2006). A simulation optimization approach for off-line blending and scheduling of oil-refinery operations. *Computers and Chemical Engineering, 30*, 614–634.

Moro, L. F. L., & Pinto, J. M. (2004). Mixed-integer programming approach for short-term crude oil scheduling. *Industrial & Engineering Chemistry Research, 43*, 85–94.

Moro, L. F. L., Zanin, A. C., & Pinto, J. M. (1998). A planning model for refinery diesel production. *Computers and Chemical Engineering, 22*, 1039–1042.

Nejad, A. T. (2007). Allocation of CO_2 emissions in joint product industries via linear programming: a refinery example. *Oil & Gas Science and Technology-Revue del Institut Francais du Petrole, 62*, 653–662.

Neiro, S., & Pinto, J. M. (2004). A general modeling framework for the operational planning of petroleum supply chains. *Computers and Chemical Engineering, 28*, 871–896.

Neiro, S. M. S., & Pinto, J. M. (2005). Multiperiod optimization for production planning of petroleum refineries. *Chemical Engineering Communication, 192*, 62–88.

Neiro, S. M. S., & Pinto, J. M. (2006). Langrangean decomposition applied to multiperiod planning of petroleum refineries under uncertainty. *Latin American Applied Research, 36*, 213–220.

Pierru, A. (2007). Allocating the CO_2 emissions of an oil refinery with Aumann-Shapley prices. *Energy Economics, 29*, 563–577.

Pinto, J. M., & Moro, L. F. L. (2000). Planning model for petroleum refineries. *Brazilian Journal of Chemical Engineering, 17*, 575–586.

Pinto, J. M., Joly, M., & Moro, L. F. L. (2000). Planning and scheduling models for refinery operations. *Computers and Chemical Engineering, 24*, 2259–2276.

Pitty, S. S., Li, W., Adhitya, A., Srinivasan, R., & Karimi, I. A. (2008). Decision support for integrated refinery supply chains – part 1. Dynamic simulation. *Computers and Chemical Engineering, 32*, 2767–2786.

Pongsakdi, A., Rangsunvigit, P., & Siemanond, K. (2006). Financial risk management in the planning of refinery operations. *International Journal of Production Economics, 103*, 64–86.

Reddy, P. C. P., Karimi, I. A., & Srinivasan, R. (2004a). A novel solution approach for optimizing scheduling crude oil operations. *AIChE Journal, 50*, 1177–1197.

Reddy, P. C. P., Karimi, I. A., & Srinivasan, R. (2004b). A new continuous time formulation for scheduling crude oil operations. *Chemical Engineering Science, 59*, 1325–1341.

Szklo, A., & Schaeffer, R. (2007). Fuel specification, energy consumption and CO_2 emission in oil refineries. *Energy, 32*, 1075–1092.

Stommel, J., & Snell, B. (2007). Consider better practices for refining operation. *Hydrocarbon Processing, 86*, 105–109.

Zhang, B. J., & Hua, B. (2007). Effective MILP model for oil refinery-wide production planning and better energy utilization. *Journal of Cleaner Production, 15*, 439–448.

Zimberg, B., & Testuri, C. E. (2006). Stochastic analysis of crude oil procurement and processing under uncertain demand for bunker fuel oil. *International Transactions in Operational Research, 13*, 387–402.

Part II
Electricity Markets and Regulation

Multivariate Modelling and Prediction of Hourly One-Day Ahead Prices at Nordpool

Jonas Andersson and Jostein Lillestøl

Abstract In this paper, we exploit multivariate and functional data techniques to capture important features concerning the time dynamics of hourly one-day ahead electricity prices at Nordpool. The methods are also used to obtain pragmatic prediction schemes for these prices. Following Huisman et al. (Energy Economics 29, 2007), we first consider the 24-hourly one-day ahead prices as a 24-dimensional variable observed on a daily basis. These are analyzed by means of a multivariate analysis of variance (MANOVA) and the results are presented by some enlightening graphs. We then account for the smoothness of the curve that these 24 values form when they are seen as a function of time of the day. This smoothness is exploited by a functional analysis of variance (FANOVA). Multivariate prediction is then addressed, first by a simple univariate ARIMA scheme combined with hourly corrections obtained from the MANOVA and then by a true multivariate state space scheme.

1 Introduction

Modelling and forecasting of electricity prices is of importance for producers as well as consumers. Both the spot market and the financial market connected to deregulated electricity markets need forecasts for capacity planning and risk management. While many features of electricity prices are in common with features of stock prices, there are also some important distinctions to make. Two such distinctions are the obvious seasonalities and the possible stationarity of electricity prices. Models containing such features are usually not preferred for stock price analysis since they clearly contradict the idea of market efficiency. There are therefore reasons not to adopt stock price models directly to electricity prices since doing so would imply ignoring important information present in historical data.

J. Andersson (✉) and J. Lillestøl
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: jonas.andersson@nhh.no; jostein.lillestol@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_9,
© Springer-Verlag Berlin Heidelberg 2010

The present paper aims to contribute to the literature about the dynamic properties of so-called spot prices in the electricity market. Our work differs from Huisman et al. (2007) in the following respects: We include a yearly component and we also suggest a variety of other methods of analysis. In particular, we show how statistical methods such as multivariate analysis of variance (MANOVA) and functional analysis of variance (FANOVA) can be used to discover such properties.

The rest of the paper is organized as follows. Section 2 briefly describes the trading mechanism on the Nordpool market and the nature of the data that we analyze. Section 3 will give an overview of some documented stylized facts of electricity prices followed by Sect. 4, which reviews the recent literature on modelling and forecasting of electricity prices. Section 5 uses MANOVA to investigate the different seasonal properties of electricity prices. In Sect. 7, it is shown how to combine results from MANOVA with a simple univariate time series technique to forecast the hourly spot prices. Section 6 shows how to exploit the smoothness properties of the daily prices profiles to reduce estimation uncertainty. This is done by means of FANOVA. In Sect. 8, methods to deal with possible volatility aspects are discussed, and in Sect. 9, a state-space approach is explored in order to study the possibility of time-varying parameters for the analysis of variance models. Section 10 concludes.

2 Nordpool and Electricity Price Data

The price data under study is system prices from the Nordpool Elspot market[1], a trading place for electricity for the Scandinavian countries, Finland and Germany. On this market, hourly contracts for physical delivery of electricity for the next 24-h period are announced once a day. The prices are formed by so-called equilibrium point trading where the market participants put their bids and offers, after which supply and demand curves of the market are determined. The intersection between these curves is the price. Bids can be posted for a particular hour (hourly bids), for a block of hours (block bids) or for producing in the hour with highest price (flexible hourly bids). The observation period of the data studied here is 1997–2007.

Congestions are treated by dividing the market into bidding areas so that the market participants must make their bids in the area where their production or consumption is physically located. The bidding areas are Sweden (SE), Finland (FI), the German area (KT), Jutland (DK2) and Zealand (DK1). Norway is divided into two or three areas depending on the capacity at a particular point in time. This division is determined by the Norwegian Transition System Operator (TSO), Statnett. As pointed out by Huisman et al. (2007), a price series of this type should be interpreted as daily observations on a 24-dimensional time series, or panel. If the prices are seen as an hourly observed time series, one would not account for the fact that the prices are actually determined once every day. From a practical point of view, the data is therefore organized according to Table 1.

[1] A more detailed description of the pricing mechanism can be found at www.nordpoolspot.com

Table 1 The organization of the data as 24 daily observed time series

	1 a.m.	2 a.m.	...	24 p.m.
01/01/97	230.18	221.27	...	220.72
01/02/97	218.27	214.79	...	228.52
01/03/97	239.62	232.93	...	240.50
01/04/97	252.34	250.41	...	248.83
01/05/97	251.31	248.54	...	254.63

2.1 Local or UTC Time

A methodological problem occurs because of daylight-saving time changes. This implies that once every spring there will be a lack of an observation at one time-point. This problem has to be handled with care, since seasonalities are of main interest here, and these are distorted by this seemingly minor problem.

By using UTC (GMT) time, we will obtain a time series that does not have this problem. However, since the peak hours of electricity prices are related to the local time that governs when electricity is more or less used, in many applications, the local time is the time to prefer. In those cases, the missing spring observation and the additional autumn observation must be aligned correctly in the dataset. The solution we have used for the missing spring observation is to impute with an interpolated value of the two surrounding ones.

2.2 Price or Log-Price?

Working with daily stock price data, a natural step before commencing with any analysis, is to compute returns by taking the first difference of the logarithms. This can, for daily data, be interpreted as a close approximation to percentage returns. The term "return" does not have the same natural interpretation for electricity spot prices since this concerns a commodity that will, in fact, be physically delivered. However, one could suspect that variation in price for high levels is larger than the variation for low levels, causing problems in many types of statistical analysis. Thereby, the logarithm of the price is a good starting point from a statistical point of view.

3 Some Stylized Facts of Electricity Prices

In this section, some initial univariate analysis of the daily means of the hourly log-prices will be made. We do this to see what stylized facts we can also expect in the multivariate setting, in order to provide realistic models. However, we will not deal with all stylized facts for the multivariate case in this paper. In practice, one would keep the modelling simple and limit the model to the most important aspects for

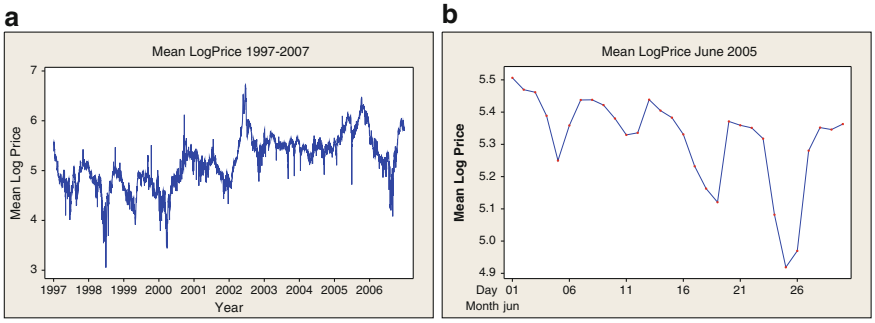


Fig. 1 Mean log-price plots. (**a**) Mean Log-Price 1997–2007, (**b**) Mean Log-Price June 2005

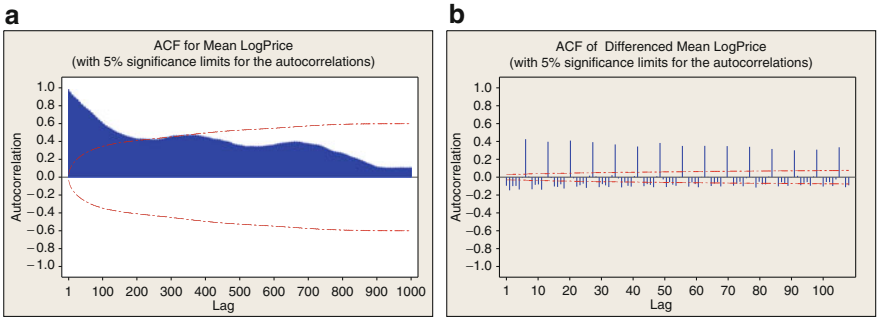


Fig. 2 Autocorrelation functions (ACF) for mean log-prices and differenced mean log-prices. (**a**) ACF for mean log-price, (**b**) ACF for first-difference of mean log-price

the given context. Figure 1 shows a time series plot of the mean log-price series, with features similar to each hourly component. Some stylized facts are discernible already from this figure. A yearly season is obviously present and there are sudden jumps of the price downwards and upwards with irregular frequency. Not as obvious, there is a slight tendency of the log-price to increase over the 11 observed years. By looking at Fig. 1b of a shorter stretch of the series, one can see that there also is an obvious weekly seasonality.

3.1 Seasonalities

Figure 2 shows the autocorrelation functions (ACF's) of the log-price and the first-difference of the log-price (which is essentially relative price changes). The seasonalities, already uncovered in the time series plots in Fig. 1, are even more visible here. The yearly variation is seen as the wave of period length approximately 365 in the ACF of the log-price. The weekly season can be seen in the ACF of the relative changes. The slow decay of these graphs combined with the fact that the

length of the cycles are equal to two natural seasonalities in nature (years) and human life (weeks), we are bound to believe that the seasonalities can be realistically modelled as deterministic.

3.2 *Stationary or Non-Stationary Models*

The question whether the log-prices are stationary or not is answered already since a deterministic seasonality has been argued to be present here. Thus, they are not assumed to be stationary. However, one outstanding issue concerning stationarity, or more precisely covariance stationarity, is the properties of the deviations from these deterministic seasonalities. For this purpose, the residuals after correction for seasonalities may be computed, by analysis of variance or regression by dummies, and studied. A time plot of such residuals is presented in Fig. 3a.

The slowly decaying autocorrelations in Fig. 3b indicate that the residual series may be non-stationary or nearly so. However, a unit root test rejects the hypothesis of non-stationarity. It turns out that the stationary ARIMA$(1, 0, 1)(1, 0, 1)_7$ process fits the series fairly well, with autoregressive coefficients of magnitude AR $= 0.914$ and seasonal AR $= 0.863$, respectively. This is not far from one, which corresponds to the non-stationary process ARIMA$(0, 1, 1)(0, 1, 1)_7$. It may of course be of some interest to conclude that the series has mean reverting features. However, if the context is short-term predictions, a long run property of this kind is of minor interest, and the simpler ARIMA$(0, 1, 1)(0, 1, 1)_7$ may be preferred. A similar discussion applies for the residual from each component series, with some indications that the stationary ARIMA$(1, 0, 1)(1, 0, 1)_7$ should be preferred for these.

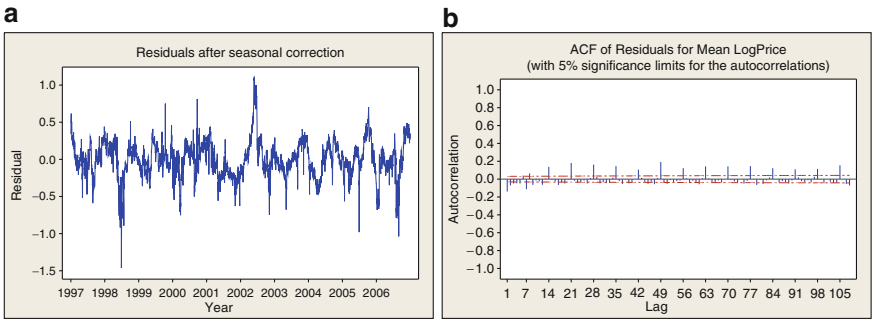


Fig. 3 Residuals from seasonally adjusted series. (**a**) Time series plot of residuals, (**b**) ACF of residuals

3.3 Distributional Assumptions

In order to see if normality assumptions may be justified we provide histograms and normal probability–probability (PP) plots of the seasonally adjusted series. In Fig. 4 we see a fairly bell-shaped distribution with somewhat heavier tails than the normal. Similar patterns exist for each of the component series. But as we shall see later, there are hourly differences. Note that since these are strongly correlated, we do not necessarily observe a strong central limit effect for the mean log-price series.

A number of methods assume normality but we see no reason to replace the error distribution here by another one, say a normal mixture distribution, which is a common way to model thick tails. In some applications, though, such as risk evaluations, it is essential to replace the assumption of normality with a more realistic one.

3.4 Jumps

The sudden outbursts, or jumps, in the price series that can be seen in Fig. 1a is a feature common with security prices, but is probably more distinct for electricity prices. This property is one of the factors making the thick tails of the distribution of price changes and should ideally be accounted for in any statistical analysis where the tails of the distributions are of relevance. This is for example, the case when risk shall be measured. One popular way of modelling jumps in the continuous finance literature is to combine a diffusion process with a marked point process, a so-called jump-diffusion process. These processes imply trajectories that usually behave quite smoothly, but suddenly with low and irregular frequency jumps up or down similar to the electricity spot price in Fig. 1a. For electricity prices, this approach is used by Weron et al. (2004).

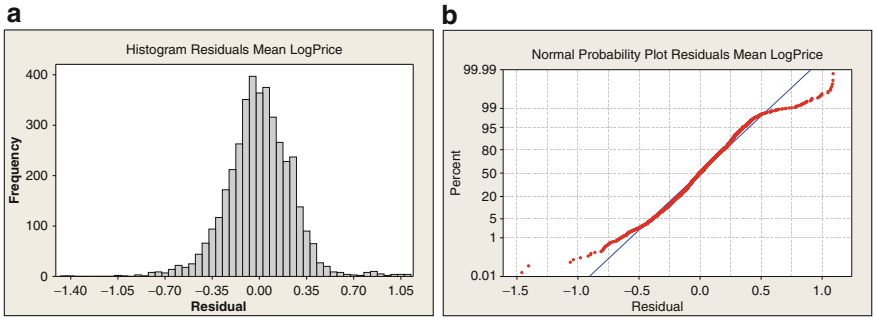


Fig. 4 Distribution of residuals from seasonally adjusted series. (**a**) Histogram of residuals, (**b**) Normal PP plot of residuals

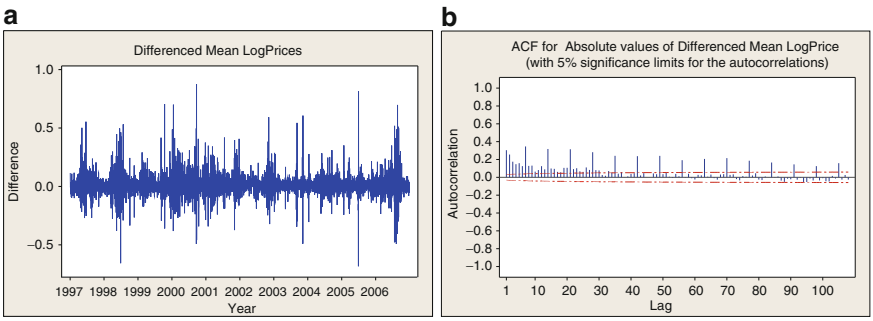


Fig. 5 Differenced mean log-prices to illustrate volatility aspects. (**a**) Time series plot, (**b**) ACF of absolute values

3.5 Volatility Clustering

Volatility clustering in the 24-vector time series of log-prices may be studied from a variety of viewpoints, either individually or in terms of a derived series that represents the day, say the daily average, median, minimum, maximum or range. We may also study the volatility directly, or in terms of the residuals after accounting for the weekly and yearly seasonalities. Note here the difference between hourly, non-vectorized volatilities and daily, cross-vector volatilities.

Volatility clustering is typically studied in terms of the differences in log-prices. As an example, Fig. 5a and 5b show the daily differenced mean log-price series and the autocorrelation function of its absolute value.

In order to explore the volatility of the mean log-price series in more detail, we have fitted an ARMAX–GARCH model to the series. On the basis of the above findings, we have specified the ARMA term as ARMA(1,1), represented the seasonalities by the covariate matrix X of seasonal dummies and specified the volatility by the common GARCH(1,1). Note that this neglects any stochastic seasonalities indicated in Sect. 3.2. It turns out that both GARCH parameters are significant with sum close to one, the case referred to as IGARCH, where the unconditional variance does not exist.

This is also representative for volatility patterns for each component series and for the residuals after accounting for the seasonalities.

Thus, there is obviously predictable volatility in electricity price changes. This could in some applications, such as risk evaluations, be readily exploited. It could also be useful to incorporate this feature when calculating prediction intervals for the prices.

4 Literature Review

This short literature survey intends to give a picture of the state-of-the-art on the topic of modelling and forecasting of electricity prices.

Deng and Jiang (2005) propose models based on Levy-driven Ornstein-Uhlenbeck processes with the main purpose of obtaining a good fit to the marginal distribution of electricity prices. Guthrie and Videbeck (2007) show that commonly used statistical models for electricity prices are unable to capture a number of properties for such series. They argue that the direct adoption of models from the security price literature is inappropriate since they cannot capture the correlation structure between different hours during the day. Hadsell and Shawky (2006) study volatility of electricity prices and its causes, such as geographical zone and transition congestion. Huisman et al. (2007) use a panel data approach to simultaneously model the day-ahead prices for all hours of the day. This is also the approach followed in the present paper. They find hourly-specific mean-reversion around an hourly-specific mean-level of the prices. They also conclude that correlation is higher between prices within different categories, such as peak-hours, than for between prices in different categories. Vehvilainen and Pyykkonen (2005) present a theory-based method where fundamentals are independently modelled after which a market equilibrium model is used to model the electricity spot price. In Walls and Zhang (2005) and Chan and Gray (2006), the authors exploit extreme value theory and the ideas of Value-at-Risk (VaR) to study electricity price risk. Weron et al. (2004) use a jump-diffusion process together with a hidden Markov model to capture the extreme spikes observed in electricity prices, while Weron and Przybylowicz (2000) investigate the property of mean-reversion in electricity prices by means of Hurst R/S analysis. Becker et al. (2007) develop a Markov switching model with transition probabilities that are allowed to depend on variables representing demand and weather.

The articles cited above did not focus mainly on the issue of forecasting. The following papers do. Conejo et al. (2005) combine a wavelet transform and ARIMA models to forecast day-ahead electricity prices. Their idea is to use the wavelet transform to obtain series more appropriate to model with ARIMA-models than the original series. Nogales and Conejo (2006) use transfer function models, while Ruibal and Mazumdar (2008) use a bid-based stochastic model to predict both hourly and average prices. This model accounts for both economic and physical features of the market. Garcia-Martos et al. (2007) use a mixed model approach where several forecasting methods are combined. There are also a number of papers exploiting artificial neural network techniques, for example, Szkuta et al. (1999) and Pao (2007).

5 Multivariate Decomposition and Modelling

We can always model log-prices at hour h day t, $y_t(h)$, as a sum of two components

$$y_t(h) = f_t(h) + e_t(h) \tag{1}$$

where $f_t(h)$ is a deterministic component representing predictable regularities, like expected log-price level and seasonalities, and a stochastic component $e_t(h)$ representing volatility and spikes, as well as serial correlation in level and volatility.

We may represent time by year (i), week (j), weekday (k) and hour (h), so that the combination (i, j, k, h) uniquely determines the observation unit. A natural decomposition of the log-prices is

$$y_{ijkh} = \lambda_h + \lambda^Y_{ih} + \lambda^W_{jh} + \lambda^D_{kh} + e_{ijkh} \tag{2}$$

with subscripts now representing time identification.

Given the one-day ahead information structure, it is natural to consider the price quotes for a given day as a 24-dimensional vector $\mathbf{y_{ijk}}$ with elements y_{ijkh}. We then have

$$\mathbf{y_{ijk}} = \boldsymbol{\lambda} + \boldsymbol{\lambda}^Y_i + \boldsymbol{\lambda}^W_j + \boldsymbol{\lambda}^D_k + \mathbf{e_{ijk}} \tag{3}$$

If we consider the more coarse classification year (i), month (j), and weekday (k), we have to add an index (l) denoting the "repeat" of the given month ($l = 4$ or 5).

$$\mathbf{y_{ijkl}} = \boldsymbol{\lambda} + \boldsymbol{\lambda}^Y_i + \boldsymbol{\lambda}^M_j + \boldsymbol{\lambda}^D_k + \mathbf{e_{ijkl}} \tag{4}$$

These representations may be used for decomposing the log-prices linearly in deterministic seasonal components without any concern about the error term. The expressions may be taken as point of departure for multivariate analysis of variance (MANOVA) modelling. However, it is not necessary that the error terms are independent, identically distributed, and definitely not multinormal as required for standard inference theory. Nevertheless, the least squares feature of MANOVA software may be used to derive a decomposition, and do exploratory data analysis, for instance, looking at the hourly profiles of the coefficients in each component. Above, we think of the categorical variables as fixed effect variables, which may be appropriate for Month and Weekday. Year may alternatively be taken as a random effect or, if we believe there is a trend, taken as a covariate.

Note also that our problem may alternatively be described within the framework of panel data modelling or multilevel statistical modelling, for which there exist software that combines the categorical features with the time series features. We will not follow this here.

We will study the representation by the year, the month and the weekday. As for the parameterization we may take one category for each of the three categorical variables as base. With 11 years of data this corresponds, including the constant term, to $1 + 10 + 11 + 6 = 28$ free parameter vectors. Alternatively, we may express it as deviations from a total mean vector, thus having $1 + 11 + 12 + 7 = 31$ parameters, but having in mind that they sum to the zero-vector over the categories for each of the three categorical variables. The estimated coefficients obtained from running a standard MANOVA program are shown in Fig. 6, where the exhibits are coefficients for the constant term (a), year (b), month (c) and weekday (d).

In exhibit (a), we see clearly the overall daily pattern with price going down from midnight to about 4 a.m., then increasing to a high at 10 a.m., and then declining with a temporal upward bump at about 6 p.m. In exhibit (b), we see the daily pattern due to year, which is fairly constant within each year. The curves for the first four years

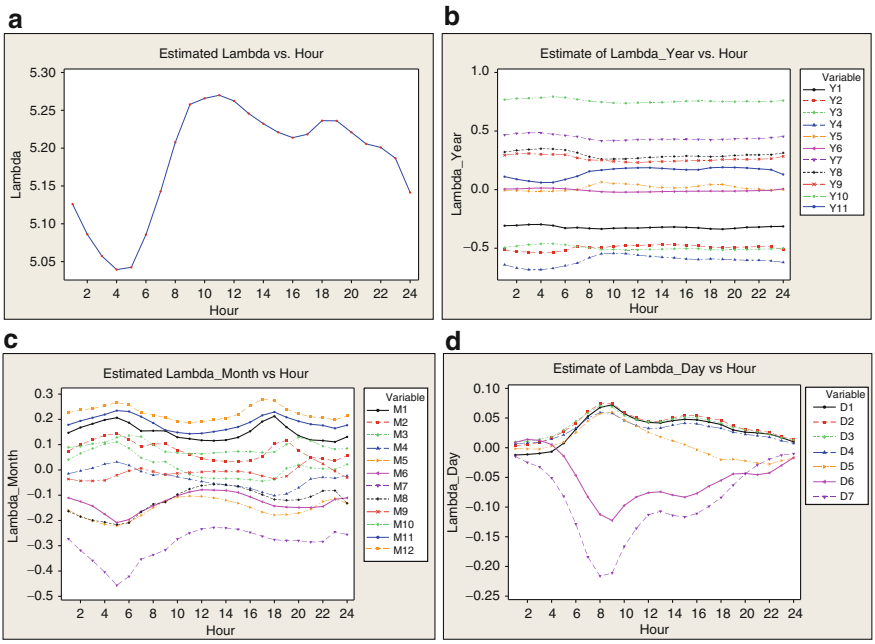


Fig. 6 MANOVA lambda parameter estimates vs. Hour. (**a**) Lambda vs. Hour, (**b**) Lambda-Year vs. Hour, (**c**) Lambda-Month vs. Hour, (**d**) Lambda-Weekday vs. Hour

(1997–2000) are below the mean level, the following (2001–2002) at the mean level, and the last five (2003–2007) are above the mean level. This may reflect an upward trend. This should not be taken as deterministic, for instance, there is a significant drop from 2006 to 2007. In exhibit (c), we see the daily pattern due to month, with the December curve on top and the July curve at the bottom, and the other monthly curves ordered in a natural fashion. Note the differences between the daily patterns between the summer and the winter months, with a downpass at night in the summer and a slight rise at night in the winter, less differences during working hours, and then more disparate again in the early evening. In exhibit (d), we see the daily pattern due to weekday, with the two weekend curves below the mean level with Sunday at the bottom, and both with a significant dip at about 8 a.m. The five ordinary working days curves are, as expected, above the mean level and fairly close together, except for Friday afternoon, which moves towards the weekend level. A closer look at these curves may reveal additional features that make practical sense.

A MANOVA also provides estimates of the standard errors of prediction. However, they are based on normality, serial independence and constant covariance matrix assumptions, which is not necessarily the case. It is therefore of interest to look into the multivariate distributional aspects of the residuals. Here we limit ourselves to exhibit the daily variation of the standard deviation, skewness and kurtosis. This is shown in Fig. 7. We see that the standard deviation is fairly stable over the

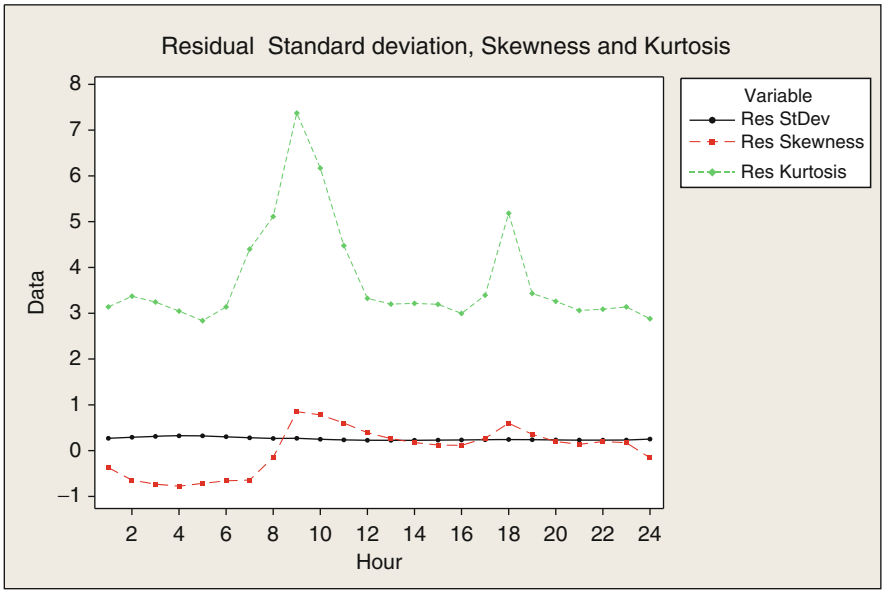


Fig. 7 Daily variation of the standard deviation, skewness and kurtosis of the residuals

day. The skewness is negative at night but then goes positive in the morning and stays positive with peaks around 9 a.m. and 6 p.m. The kurtosis is slightly above 3, the level corresponding to normal, with a strong peak above 7 around 9 a.m. and a smaller peak around 6 p.m.

The MANOVA-model may also be written on multivariate regression form, that is,

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{B} + \boldsymbol{E} \tag{5}$$

where $\mathbf{Y}$ and $\mathbf{E}$ are $T \times 24$ matrices and $\mathbf{X}$ is the $T \times 28$ design matrix and $\mathbf{B}$ is the 28×24 matrix of regression coefficients, with T being the total number of observations. In our case T is 4,017.

The design matrix will consist of 0s and 1s in the pattern that defines the calendar structure of the problem. If we have chosen the alternative parameterization with parameter vectors summing to one for each categorical calendar variable, the left out category for each type (usually the last) will be represented by -1, in order for the regression parameters of the type to sum to the zero vector. MANOVA software typically provides the estimated coefficients only as option, and it is important to know the kind of parameterization used. Some software also provides the generated design matrix, which may be used for later regression purposes (and also tells the kind of MANOVA parameterization used).

In the above, we have set aside the possible time series features of the series. Within the given framework this can be studied by looking at the residuals after the MANOVA estimation (or the multivariate regression estimation). For the current data they turned out to have features like a multivariate IMA(1) series. This may

indicate the alternative modelling approach, where we model a multivariate time series with the calendar variables as covariates. However, multivariate time series software may not be readily available, and are most often limited to vector autoregressive time series, possibly with covariates. One notable exception is SCA, which has a framework for identification and estimation of VARIMA models.

6 Functional Data Approach

A feature not taken into account in the MANOVA in Sect. 5 is the fact that the co-variation of the different spot prices for each day is not totally arbitrary. There is a smoothness in the price profiles, that is, the spot price seen as a curve over the hours of one day. While maintaining the interpretation of the MANOVA, this smoothness can be accounted for and thereby can remove some of the estimation uncertainty. In this section, this will be done by means of functional data analysis, see, for example, Ramsay and Silverman (1997) for an excellent review of this topic. Here a very brief introduction will be given to enable us to describe the technique of interest in this paper, functional analysis of variance (FANOVA).

6.1 Notation

A functional data observation can be represented by a vector

$$\boldsymbol{y}_t = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{6}$$

where each of the n points are observations from a function, $y_t(h)$, with some assumed smoothness properties, for example, continuous second-order derivatives. Each point in the vector $\boldsymbol{y}_t$ is observed for the argument value h_r, $r = 1, 2, ..., n$, in this paper the same values for all observations. n would in the present analysis be the number of observations per day, 24. The set of T observations of the variable is often represented in terms of the matrix

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{T1} & y_{T2} & \cdots & y_{Tn} \end{pmatrix} \tag{7}$$

For the present, dataset T is 4,017, that is, the number of days observed.

6.2 System of Basis Functions

A common way to impose the assumption of smoothness, which will also be used in this paper, is to choose an appropriate system of P basis functions, $\boldsymbol{\phi}(h) = [\phi_1(h), \phi_2(h), ..., \phi_P(h)]'$ that can capture the features of the observed data. If well choosen, a linear combination of these functions can approximate the observed data sufficiently well, even if P is choosen significantly smaller than n. In the present analysis, the basis will be sine and cosine functions because of the periodic nature of our data. Our functions can now be represented by

$$y_t(h) = \boldsymbol{c}_t'\boldsymbol{\phi}(h) \tag{8}$$

where the $P \times 1$-vector $\boldsymbol{c}_t$ contains coefficients determining each data point. The entire dataset is often represented by

$$Y = C\Phi \tag{9}$$

where

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1P} \\ c_{21} & c_{22} & \cdots & c_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ c_{T1} & c_{T2} & \cdots & c_{TP} \end{pmatrix} \tag{10}$$

and

$$\Phi = \begin{pmatrix} \phi_1(h_1) & \phi_1(h_2) & \cdots & \phi_1(h_r) \\ \phi_2(h_1) & \phi_2(h_2) & \cdots & \phi_2(h_r) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_P(h_1) & \phi_P(h_2) & \cdots & \phi_P(h_r) \end{pmatrix} \tag{11}$$

The question of how to calculate the C-matrix will be answered differently depending on which basis system has been chosen. The perhaps obvious solution, the best mean square fit $Y\Phi'(\Phi\Phi')^{-1}$ is sometimes appropriate.

In the present case, however, we are using a Fourier basis, defined by $\phi_0(h) = 1$, $\phi_{2r-1} = sin(r\omega h)$ and $\phi_{2r} = cos(r\omega h)$ where $\omega = 2\pi/H$ and H is the length of the observation interval (24 in the present case). It is well known that the coefficients in this case are best calculated by means of the fast Fourier transform (FFT). P is here set to 10. Loosely speaking, the reduction of the multivariate analysis of variance problem has been reduced from 24 to 10. A more elaborate analysis, not done here, would be to use a roughness-penalty approach in the estimation in order to smooth the function even more. The smoothness of the estimated function, in the present analysis totally governed by the chosen number of basis functions, could then be determined by the observed data, that is, through cross-validation.

6.3 Functional Analysis of Variance (FANOVA)

The functional model we will use is written as

$$y_{ijkl}(h) = \lambda(h) + \lambda_i^Y(h) + \lambda_j^M(h) + \lambda_k^D(h) + e_{ijkl}(h) \tag{12}$$

where notation analogous to the one in Sect. 5 is used. The estimation is made by the package *fda*-package (Ramsay et al., 2007) in the statistical programme package *R* (R Development Core Team, 2007).

Just as for the MANOVA in Sect. 5, which can be rewritten according to (5), the model (12) can be formulated as a functional regression model. In this model, only the dependent variable is functional. The explanatory variables are represented by the same matrix X as in the MANOVA.

$$y_t(h) = \boldsymbol{x}_t \boldsymbol{\lambda}(h) + e_t(h) \tag{13}$$

where $\boldsymbol{x}_t$ is the t'th row of the X-matrix, $y_t(h)$ is the function $y_{ijkl}(h)$ and t now is an index which uniquely defines a combination of the quadruple (i, j, k, l). $\boldsymbol{\lambda}$ is a 28-dimensional vector of functions representing the grand-mean function of prices over one day, 6 functions representing the functions specific for weekdays, 11 month-specific functions and 10 year-specific functions. Since the estimation is performed in such a way that all weekday-functions should sum to zero, the 7th weekday-function is obtained by just negating the sum of the 6 estimated functions. Analogous reasoning produces the remaining functions for the month- and year-specific effects.

The estimation is performed by representing $y_t(h)$ and the λ-functions in the Fourier-basis. This results in a multivariate regression where the coefficients of the observed y-function are regressed on X yielding the coefficients of the λ-functions, $\boldsymbol{b}$. From these, the λ-functions can be recovered by $\lambda(h) = \boldsymbol{b}'\boldsymbol{\phi}(\boldsymbol{h})$. See Ramsay and Silverman (1997), Chap. 9 for a more comprehensive presentation of FANOVA. The results of the analysis can be seen in Fig. 8.

The pattern is consistent with the MANOVA. The improvement of the method consists of less estimation variability.

7 Prediction via Univariate ARIMA Modelling

An alternative to VARIMA predictions with covariates may be to model the daily mean log-prices as a univariate time series, and get predictions for the individual hours by correction factors estimated from daily profiles, which may depend on the calendar. In the time series of daily averages, we still have two seasons, the weekly and the yearly. For market predictions locally, the weekly season is of prime importance, and will enter in the ARIMA specification, while the yearly season can be

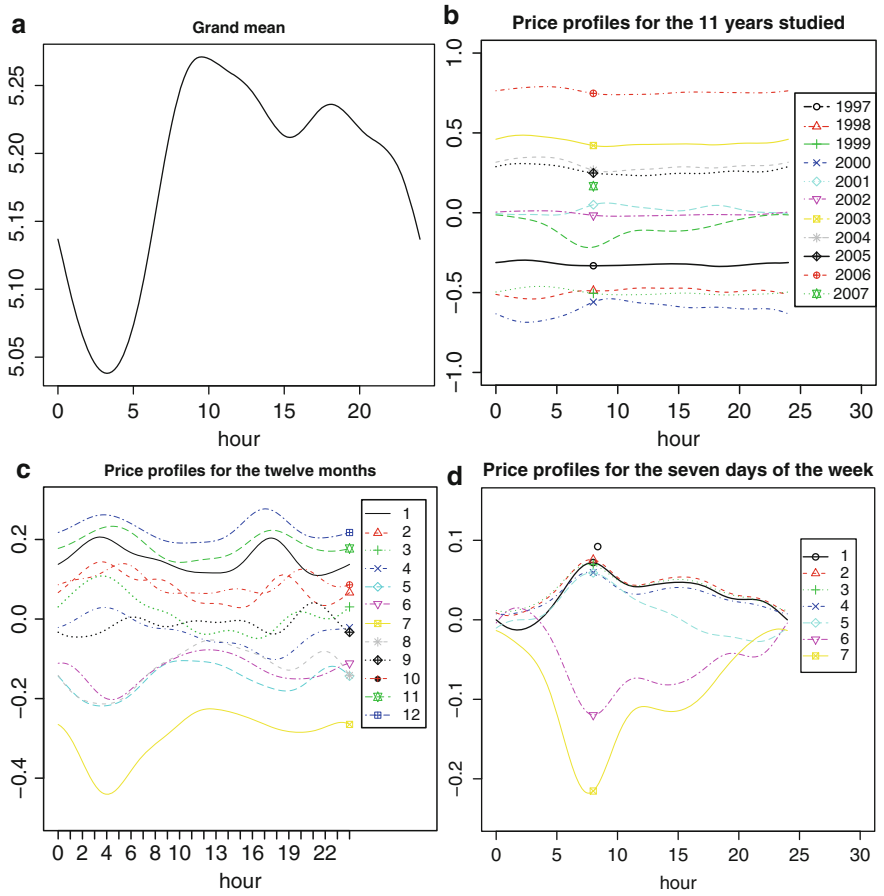


Fig. 8 FANOVA lambda function estimates vs. Hour. (**a**) Lambda vs. Hour, (**b**) Lambda-Year vs. Hour, (**c**) Lambda-Month vs. Hour, (**d**) Lambda-Weekday vs. Hour

dealt with in a variety of ways, among others (a) take as a covariate (b) performing a local fit or (c) ignore it. The latter may be in order as a pragmatic choice, since the likely model involves differencing, so that the predictions will depart from the local level anyway.

Taking VARIMA$(0, 1, 1)(0, 1, 1)_7$ with monthly covariates as a preferred model for the 24-dimensional log-price vector, it will be consistent to take a univariate ARIMA$(0, 1, 1)(0, 1, 1)_7$ with monthly covariates as the preferred model for the daily average log-prices. This is not so when autoregressive terms are added, or if a VAR-approach is taken.

In fact, autocorrelation plots for the mean-log-price reveal that differencing and seasonal differencing are necessary to achieve stationarity. In Fig. 9, we show the

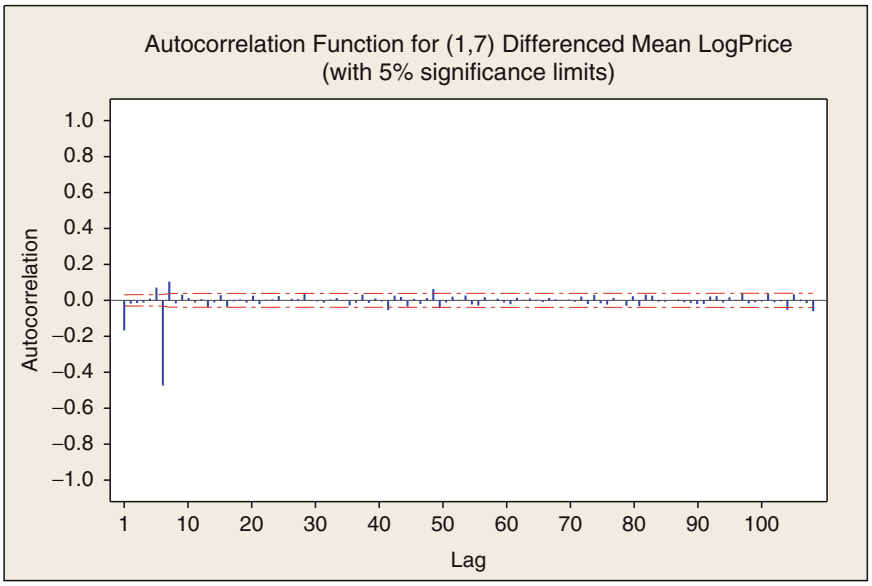


Fig. 9 Autocorrelation of (1,7)-differenced log-price

autocorrelation plot after this differencing, consistent with an MA(1) and an MA(7) component, with the common side lobes at the seasonal lag.

The estimated model turned out as follows:

```
ARIMA Model: MeanLog-Price

 Type       Coef  SE Coef       T      P
 MA  1   0.1818   0.0155   11.70  0.000
 SMA 7   0.8887   0.0074  120.57  0.000

Residuals:     SS =  24.1882
               MS =  0.0060  DF = 4007

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

 Lag             12     24     36     48
 Chi-Square    89.2  117.6  140.5  160.8
 DF              10     22     34     46
```

The residuals from the ARIMA fit indicate possibilities for improvement, but no natural simple alternative model seems at hand. This is probably due to some structural changes in a series of this length, not accounted for in the strict ARIMA context. If we estimate this model, using shorter stretches of data, say 2–3 months, the fit is generally very good, residual variance lower, with the seasonal MA(7) still highly significant, while the MA(1) term is significant in most cases, although not in the case of the last 2–3 months of the data from 2007.

Table 2 Multiplicative hourly factors to daily mean price predictions. Numbers based on MANOVA

Hours 01–06	0.9442	0.9073	0.8816	0.8660	0.8685	0.9070
Hours 07–12	0.9603	1.0246	1.0771	1.0858	1.0903	1.0822
Hours 13–18	1.0643	1.0500	1.0382	1.0309	1.0354	1.0543
Hours 19–24	1.0540	1.0382	1.0224	1.0177	1.0032	0.9588

Table 3 Multiplicative hourly factors to daily mean price predictions. Numbers based on FANOVA

Hours 01–06	0.9093	0.8775	0.8637	0.8684	0.8940	0.9423
Hours 07–12	1.0048	1.0593	1.0864	1.0880	1.0803	1.0721
Hours 13–18	1.0597	1.0415	1.0280	1.0299	1.0433	1.0519
Hours 19–24	1.0462	1.0335	1.0242	1.0146	0.9919	0.9525

With z_t being the mean log-price at day t, the one-day ahead prediction at time u is

$$\hat{z}_u(1) = z_u + z_{u-6} - z_{u-7} - \theta a_u - \theta_7 a_{u-6} + \theta\theta_7 a_{u-7}$$

where θ and θ_7 are the estimated moving average coefficients and $a_t = z_t - \hat{z}_{t-1}(1)$, that is, past prediction errors. From these predictions, daily prices can be obtained by exponentiation. Predictions for each hour may then be found through multiplications by the hourly factors from the MANOVA- or FANOVA-analysis given in Tables 2 and 3, respectively. It is worthwhile to note that the one-step ahead prediction for our model also can be represented by two intertwined exponentially weighted moving average schemes.

8 Multivariate Volatility

Until now, our main concern has been multivariate modelling and the prediction of price level. The implicit assumption has been constant residual covariance matrix. This may not be the case, and there are two aspects to this: (a) volatility depending on the seasons and (b) conditional volatility. In Sect. 3.5, we have seen that mean log-prices exhibit univariate volatility clustering. This feature should therefore be expected in the multivariate context as well. The concern for volatility may be the prime concern in some financial contexts, but just nuisance in the operational context. To get improved point predictions of prices it seldom helps to add all data features to the model, although this may provide a more realistic uncertainty evaluation. We will briefly look into the two volatility aspects.

To get some insight to how the volatility depends on the season we have taken the log-squared residuals from the MANOVA analysis of Sect. 5, and performed another MANOVA, and plotted daily curves for each effect, main effect, year, month and weekday, similar to Fig. 6. The main conclusions from these plots (not shown) are: (a) Higher volatility at night (b) different levels between years (c) higher volatility

in July/August and December/January but lower in March and (d) all days about equal and constant effects, except Sundays which is both higher and peaked in early morning.

Now to the volatility clustering aspect: Multivariate volatility models are mainly studied in connection with volatility in returns, and where volatility is the main object of study. There are many proposals of such models in the literature, mostly of the GARCH type, see Bauwens et al. (2006) for a review. A problem with such models is the large number of parameters to be estimated, in particular, when they come in addition to the level parameters, as is the case here. In practice, it is often sufficient to limit attention to models of order (1,1). The curse of dimensionality may also be solved by index models, where each component is expressed in terms of a representative index series. In a sense, this is what we do when we have used the daily mean log-price as the basis and scaled the result according to an hourly index. If we model this representative series both with respect to time varying and conditional level and volatility, this will imply a joint scaling, convenient but not necessarily realistic.

Now consider the modelling of the conditional (residual) covariance matrix Σ_t written as

$$\Sigma_t = S_t C_t S_t \tag{14}$$

where S_t is a diagonal matrix of conditional standard deviations at time t and C_t is the correlation matrix. A possible modelling approach is to take component volatility as univariate GARCH and specify a model for the correlations. A class of models is the dynamic correlation models, where C_t is expressed by an updating scheme in terms of C_{t-1} and a correlation matrix based on recent observations, see Tsay (2006) for some suggestions.We will not go into further detail here, but note that the dynamic prediction approach presented in the next section involves updating of covariance matrices.

9 Multivariate Dynamic Modelling

In practice, one may want to make one-day ahead predictions of the hourly (log)prices based on an online updating scheme adaptive to possible parameter changes over time. One possibility is to use standard recursive univariate forecasting procedures to the mean log-prices, and then use hourly correction factors, for instance, using multiplicative factors after exponentiation, as we did above for the univariate ARIMA predictions. Among readily available procedures are exponential smoothing schemes (Holt–Winter) and state space model schemes (Kalman). Similar multivariate schemes exist, but may not be readily available as canned software.

A natural framework for dynamic modelling is to consider the 1×24 dimensional observation vector $\boldsymbol{y}_t$ expressed by

$$\boldsymbol{y}_t = \boldsymbol{x}_t \cdot \boldsymbol{B}_t + \boldsymbol{e}_t \tag{15}$$

together with the state equation for the 28×24 dimensional matrix $\boldsymbol{B}_t$

$$\boldsymbol{B}_t = \boldsymbol{B}_{t-1} + \boldsymbol{D}_t \tag{16}$$

Note that the observation equation is just line t of our previous regression equation $\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{B} + \boldsymbol{E}$ in terms of all observations, but now with the 28×24 dimensional regression matrix $\boldsymbol{B}_t$ depending on t.

Here we assume that the error terms $\boldsymbol{e}_t$ and $\boldsymbol{D}_t$ are both independent over time and mutually independent. For the between-hour covariances, we assume

$$cov(e_{ti}, e_{tj}) = \sigma_{ij} \tag{17}$$

$$cov(D_{ti}, D_{tj}) = \sigma_{ij} \cdot \boldsymbol{W} \tag{18}$$

where $\boldsymbol{W}$ is a known 28×28 matrix. Note that intraday correlation structure given by $\Sigma = (\sigma_{ij})$ is assumed common to both equations.

This model fits into the framework of multivariate dynamic linear models (MDLM) as described in West and Harrison (1997), which also allows a more extensive state equation with cross-dependencies expressed by premultiplication of the lagged state variable by a time-dependent matrix G_t, as well as having known time-dependent multiplicative factors to the covariances. With distributional assumptions on the error terms as well as prior distributions on the parameters $\boldsymbol{B}_0$ and $\boldsymbol{\Sigma}$, a Bayesian updating scheme is established, which provides the posterior distribution, given data, for $(\boldsymbol{B}_t, \boldsymbol{\Sigma})$, as well as the predictive distribution of $\boldsymbol{y}_t$. Although the distributions involved are fairly complicated, involving the matrix normal/inverse Wishart distribution, the recursive scheme for the current parameters is easily programmed. The iterations need starting values as well as specification of the matrix $\boldsymbol{W}$. If $\boldsymbol{I}_t$ denotes the observational information up to time t, the scheme goes as follows ($t = 1, 2, ...$)

1. Posterior distribution of $(\boldsymbol{B}_{t-1}, \boldsymbol{\Sigma})$ given $\boldsymbol{I}_{t-1}$.
2. Prior distribution of $(\boldsymbol{B}_t, \boldsymbol{\Sigma})$ given $\boldsymbol{I}_{t-1}$.
3. Predictive distribution of $\boldsymbol{y}_t$ given $\boldsymbol{I}_{t-1}$.
1. Posterior distribution of $(\boldsymbol{B}_t, \boldsymbol{\Sigma})$ given $\boldsymbol{I}_t$.

$\cdots$ Repeating for each t.

Although this scheme is derived under very specific distributional assumptions, not likely to be fulfilled in reality, the predictive abilities may compare well with alternatives, and may be preferred for its adaptive and recursive features. We will here gain some insight by looking at the predictions of the log-prices for the week December 10–16, 2007, in comparison with the observed log-prices this week and the week before, as seen in Fig. 10.

We see the common hourly pattern of log-prices in the predictions as well as in the observed series, but the daily curves are not ordered in the expected pattern presented earlier. Consider first the prediction in exhibit (a). Here Monday is the lowest curve, and the expected lowest, Saturday and Sunday, are in the middle. We also see a spike on Tuesday 18 h, and a corresponding low at the same time

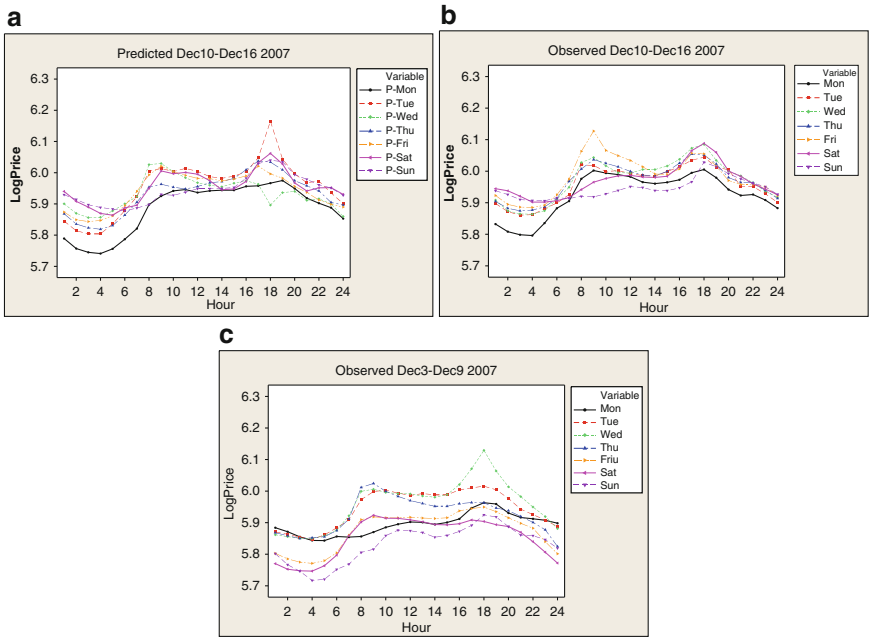


Fig. 10 Predictions Dec10-Dec16 2007. (**a**) Predicted, (**b**) Observed, (**c**) Observed week before

the following Wednesday. If we look at the observed log-prices in exhibit (b), we see good agreement with the predictions, better in the beginning of the week than at the end. We also see a spike on Friday at 9 h that was not predicted. For a part explanation of these findings, we may look at the observed pattern the week before in exhibit (c), where Monday was lower than expected. Combined with the usual low Sunday, this has lead to lowered overall predictions for the Monday of the coming week. The predictions for the prior week (not shown), have the common pattern with Saturday and Sunday as the low curves, but with the outliers at 18 h on Tuesday and Wednesday. This represents some experience from the recent past, still influencing predictions, but is not observed in the current two weeks.

10 Conclusion

We have reviewed some stylized facts of electricity prices and illustrated, on Nordpool Elspot data from 1997 to 2007, how to discover them with multivariate and functional data techniques. From an exploratory point of view, the methods produce graphs which describe the daily price profiles for different categories of the data, such as one particular month, thereby enabling comparisons between categories. It is also shown how these techniques can be used together with simple univariate

forecasting methods to predict the entire daily price profile of electricity spot prices. Furthermore, a multivariate dynamic linear model, enabling time-varying parameters, is described and illustrated on the Nordpool data. The need for allowing for the seasonalities studied in this paper is apparent. However, if used in forecasting for trading purposes, this is probably rather a requirement for not loosing money than a possibility for abnormal returns.

Acknowledgements The authors are grateful to Fridthjof Ollmar and two anonymous referees for helpful comments and to Nordpool for providing us with the data.

References

Bauwens, L., Laurent, S., & Rombouts, J. (2006). "Multivariate GARCH models: a survey". *Journal of Applied Econometrics*, *21*, 79–109.

Becker, R., Hurn, S., & Pavlov, V. (2007). "Modelling spikes in electricity prices". *Economic Record*, *83*, 371–382.

Chan, K., & Gray, P. (2006). "Using extreme value theory to measure value-at-risk for daily electricity spot prices". *International Journal of Forecasting*, 22.

Conejo, A. J., Plazas, M. A., Espinola, R., & Molina, A. B. (2005). "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models". *IEEE Transactions on Power Systems*, *20*, 1035–1042.

Deng, S. J., & Jiang, W. J. (2005). "Levy process-driven mean-reverting electricity price model: the marginal distribution analysis". *Decision Support Systems*, *40*, 483–494.

Garcia-Martos, C., Rodriguez, J., & Sanchez, M. J. (2007). "Mixed models for short-run forecasting of electricity prices: Application for the Spanish market". *IEEE Transactions on Power Systems*, *22*, 544–552.

Guthrie, G., & Videbeck, S. (2007). "Electricity spot price dynamics: Beyond financial models". *Energy Policy*, *35*, 5614–5621.

Hadsell, L., & Shawky, H. A. (2006). "Electricity price volatility and the marginal cost of congestion: an empirical study of peak hours on the NYISO market, 2001–2004". *Energy Journal*, *27*, 157–179.

Huisman, R., Huurman, C., & Mahieu, R. (2007). "Hourly electricity prices in day-ahead markets". *Energy Economics*, 29.

Nogales, F. J., & Conejo, A. J. (2006). "Electricity price forecasting through transfer function models". *Journal of the Operational Research Society*, *57*, 350–356.

Pao, H. T. (2007). "A neural network approach to m-daily-ahead electricity price prediction". In *Advances in Neural Networks - ISNN 2006, Pt 2, Proceedings* (vol. 3972, pp. 1284–1289) *Lecture Notes in Computer Science*.

R Development Core Team (2007). *R: a language and environment for statistical computing*, R foundation for statistical computing, Vienna, Austria, ISBN 3-900051-07-0.

Ramsay, J., & Silverman, B. (1997). *Functional data analysis*. Springer.

Ramsay, J. O., Wickham, H., & Graves, S. (2007). *fda: Functional data analysis*.

Ruibal, C. M., & Mazumdar, M. (2008). "Forecasting the mean and the variance of electricity prices in deregulated markets." *IEEE Transactions on Power Systems*, *23*, 25–32.

Szkuta, B. R., Sanabria, L. A., & Dillon, T. S. (1999). "Electricity price short-term forecasting using artificial neural networks." *IEEE Transactions on Power Systems*, *14*, 851–857.

Tsay, R. (2006). "Multivariate volatility models". In H. Ho, C. Ing, & T. Lai (Eds.), *Time series and related topics: in memory of Ching-Zong Wei* (Vol. 52, pp. 210–220) *IMS Lecture notes-monograph series*.

Vehvilainen, L., & Pyykkonen, T. (2005). "Stochastic factor model for electricity spot price - the case of the Nordic market". *Energy Economics*, *27*, 351–367.
Walls, W. D., & Zhang, W. (2005). "Using extreme value theory to model electricity price risk with an application to the Alberta power market". *Energy Exploration and Exploitation*, *23*, 375–403.
Weron, R., Bierbrauer, M., & Truck, S. (2004). "Modeling electricity prices: jump diffusion and regime switching". *Physica a-Statistical Mechanics and Its Applications*, *336*, 39–48.
Weron, R., & Przybylowicz, B. (2000). "Hurst analysis of electricity price dynamics". *Physica A*, *283*, 462–468.
West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models*. Springer.

Appendix: Abbreviations

$ARIMA(p, d, q)(P, D, Q)_s$: Seasonal Integrated Autoregressive Moving Average model with p autoregressive terms, d differences, q moving average terms, P seasonal autoregressive terms, D seasonal differences, Q seasonal moving average terms and seasonality s.
$GARCH(q, p)$: Generalized Autoregressive Conditionally Heteroskedastic model with q lagged squared error terms and p lagged conditional variances.
$IGARCH(q, p)$: A $GARCH(q, p)$ model where the sum of the $p + q$ coefficients equal one.
ARMAX: An $ARIMA(p, 0, q)$-model with added explanatory variables.

Time Regularities in the Nordic Power Market: Potentials for Profitable Investments and Trading Strategies?

Ole Gjølberg

Abstract Electricity is a nonstorable commodity. Consequently, electricity prices will follow fairly regular fluctuations in demand, stemming from time dependent variations in economic activity and weather conditions. However, it is possible to store electricity as a different energy carrier (e.g., hot water) and both consumers and producers have some leeway for changing behavior in order to take advantage of price regularities. Thus, the price regularities should be within arbitrage limits, and one would expect price regularities to be reduced over time as a result of investments that increase flexibility in consumption as well as production. In this article, hourly, daily, and weekly prices and price changes at the Nordic power exchange (Nord Pool) are analyzed over the period January 1995–December 2006. The tentative conclusion from the statistical analysis is that the price regularities may offer potentials for profitable investments in flexibility as well as profitable trading strategies.

1 Introduction

In well functioning commodity markets, excessive spatial price differentials will quickly be removed through arbitrage, so that prices for identical goods at different locations do not persistently differ by more than transportation and transaction costs incurred by moving goods from one location to another. Likewise, price changes over time for a commodity will not systematically and persistently move outside a range defined by the cost-of-carry, i.e., capital cost (including a risk premium), storage costs, insurance, etc. Neither will there be persistent regularities in price movements that can be utilized in order to make abnormal profits.

O. Gjølberg
Department of Economics and Resource Management, UMB, Taarnbygningen, 1432 Aas, Norway
and
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: ole.gjolberg@umb.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_10,
© Springer-Verlag Berlin Heidelberg 2010

Spatial and intertemporal arbitrage typically requires goods to be storable. In this article, we study intertemporal price regularities for electric power. Once produced, electricity as such is nonstorable. However, electricity may be stored as a different energy carrier. Prior to production, electricity may be stored as, e.g., a water reservoir or as a tank of oil or a pile of coal. After being produced, electric power may be stored as hot water or the electricity may be used to lift water upstream for later use. Furthermore, electric power may be stored as hydrogen produced through electrolyses.

Although possible, such storage of electric power is confronted by several technological and physical restrictions. While electricity demand changes in a fairly predictable way over the week or from one season to another, it is not easy to adjust supply accordingly. There is a limited flexibility in most electricity generating technologies. However, contrary to production at a nuclear plant, scheduled hydro-based electricity production can to some extent be adjusted to expected short-term variations in demand. But also for hydro power producers, there are limits to flexibility and there are costs involved in trying to scale supply according to demand variations. The laws of thermo dynamics apply and in order to, for instance, take advantage of regular price increases by using the water twice the producer must calculate the costs involved.

Consumers may also to some extent adjust their planned consumption according to intertemporal price regularities. Other energy carriers can be used as a substitute for electricity in order to heat water and buildings when the electricity price is temporarily high. This requires technological flexibility through, for instance, investments in dual-fuel type of boilers. For consumers to consider such investments, the market must function in such a way that it is practically possible to take advantage of price variations. This means that technologies must be installed that makes it possible for consumers to pay electricity consumption according to day of the week and even time of the day, known as real-time-pricing (RTP).[1] The introduction of such technologies on a larger scale than we have seen so far will most likely contribute to changes in demand patterns over time and in that way smooth price movements.

2 Literature on Price Relationships in the Power Markets

While there is a huge body of empirical studies on most commodity markets, covering topics such as market efficiency, arbitrage, spot–futures–storage relationships, etc., the literature on price relationships in the electricity markets is more modest. The major reason for this is obviously related to the fact that the electricity markets until some 10–15 years ago were tightly regulated in more or less all countries worldwide. However, the number of electricity market studies has increased rapidly

[1] For a survey on research on RTP in electricity markets, see Borenstein (2009).

since deregulation, in which the Nordic countries and the Nordic power exchange (Nord Pool) have played a leading role, subsequently followed up by the European energy exchanges (EEX and APX).

A seminal paper on the spot–forward price relationship for electricity based on data from Nord Pool was presented by Lucia and Schwartz (2002). The paper examines the regular patterns in the behavior of electricity prices and the shape of the futures/forward curve. Byström (2003) discusses hedging performance using the futures contracts at Nord Pool, while Wilkens and Wimschulte (2007) and Shawky et al. (2003) investigate the pricing of electricity futures at the EEX and in the U.S., respectively. Gjølberg and Johnsen (2001) discuss the relationship between storage volumes in terms of water reservoir fillings and the Nord Pool spot/futures spread. Koekebakker and Ollmar (2005) present empirical evidence on the forward curve at Nord Pool, while Kristiansen (2007) focuses on the monthly forward contract, also at Nord Pool. Cartea and Villaplana (2008) analyze the seasonal pattern of forward prices in the Nordic as well as the English, Welsh and U.S. markets. Outside the Nordic/European power markets, Bessembinder and Lemmon (2002) discuss optimal hedging in the electricity applying data from the Pennsylvania, New Jersey, Maryland (PJM), and the California Power Exchange (CALPX). Longstaff and Wang (2004) conduct an empirical analysis of forward pricing at the PJM market, while Woo et al. (2001) analyze cross hedging and forward-contract pricing of electricity in the U.S. market.

As regards spot price behavior in the electricity markets, a series of studies has emerged during recent years. Weron (2000) analyzes price volatility while Simonsen et al. (2004) and Simonsen (2005) present "the stylized facts" of the pricing, i.e., seasonality, return distributions, volatility clustering, and mean reversion at Nord Pool. Burger et al. (2004) present a model for spot pricing at EEX, taking into account, among other factors, seasonal patterns and mean reversion. Robinson and Baniak (2002) study the volatility of spot prices in the English and Welsh electricity pool, suggesting that generators with market power may have an incentive to create volatility in the spot market in order to benefit from higher risk premia in the contract market. Worthington et al. (2005) analyze spatial transmission of price volatility in five regional Australian electricity markets, while Park et al. (2006) conduct a similar analysis of spatial price relationships across different regional electricity markets in the U.S. Yet another category of papers aim at making forecasting models for different power markets (e.g., Gareta et al. 2005; Yamin et al. 2004; Conejo et al. 2005).

The present article is a contribution to the understanding of spot price behavior in one of the most advanced and active power markets over the last 15–20 years, i.e., the Nordic power market. Based on price observations from Nord Pool[2] every hour, 24 h a day over the period 1 January 1995 to 31 December 2006 (4,383 days; 105,192 h), a set of distinct time regularities is revealed and discussed: "The-day-of-the-week pattern," "The week-end pattern," "The time-of-the-day pattern," and

[2] The prices are the hourly system prices at the Nordic power exchange Nord Pool. Data were downloaded from Nord Pool's server.

"The Mean reversion pattern." We describe these patterns and we search for possible changes in patterns over time. Finally, we indicate to what extent there are possible arbitrage and investment opportunities based on the observed price behavior.

3 Day-of-the-Week Pattern

As for other power markets, the price in the Nordic power market is extremely volatile. Figures 1 and 2 describe the weekly (mean) price and the weekly percentage price changes over the period January 1995–December 2006. As can be seen, the price fluctuated heavily around a mean of approximately NOK 146/MW during the period 1995 through the summer of 2003. Then, a temporary extreme surge introduced a period of higher prices towards 2007. Figure 2 visualizes the extreme volatility in this market. Disregarding the spikes (normally the result of extreme weather conditions), more than 25% of the 652 weeks had a price change of $\pm 10\%$ or more. The weekly standard deviation is 12.5%, annualized close to 90%. This is substantially higher than the volatility in the oil market during the same period (roughly 30% on an annual basis) and very much higher than the stock market volatility (20–24%).

The weekly means smoothen out the day-to-day variations (as well as the intraday variations). Thus, there is a highly regular pattern in price levels and price changes through the week. Figure 3 describes this pattern, i.e., high Monday prices, slightly lower on Tuesday, Wednesday, and Thursday. Comes Friday and the price is

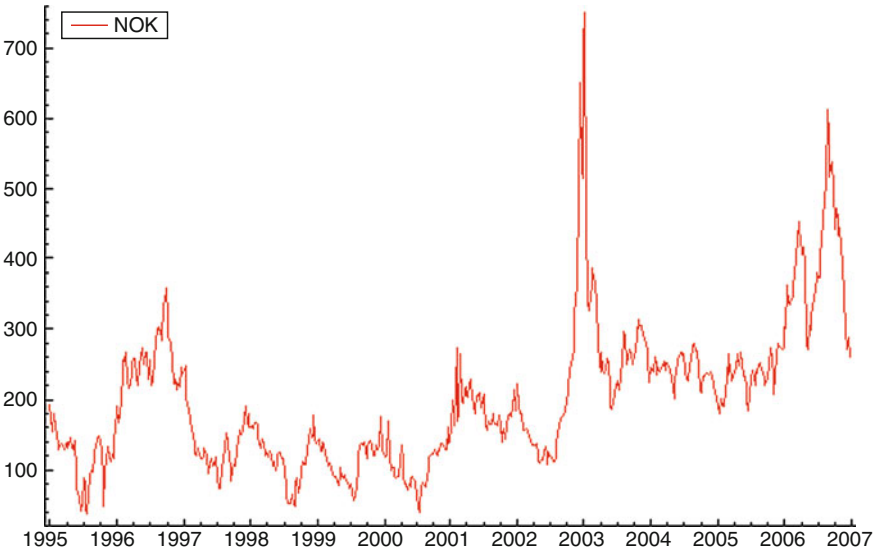


Fig. 1 Weekly mean prices 1995–2006, NOK/MWh

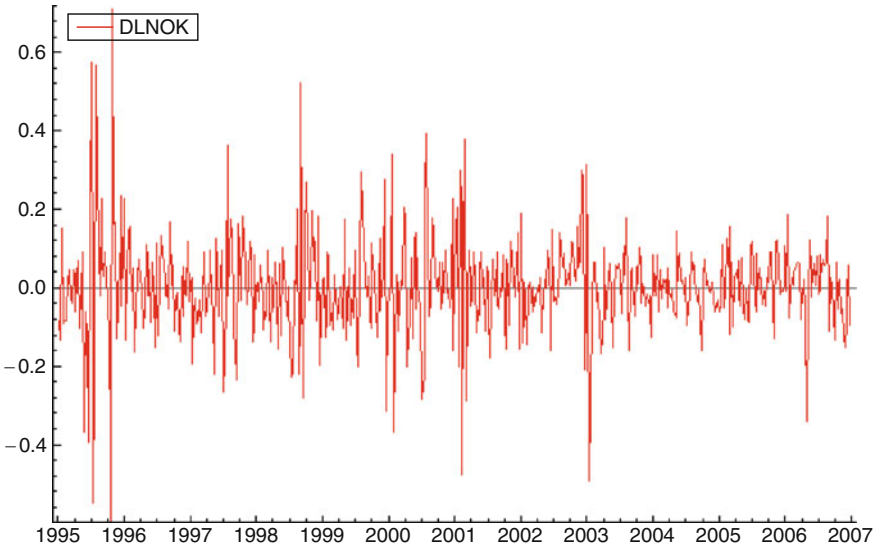


Fig. 2 Percentage price changes, weekly means, 1995–2006

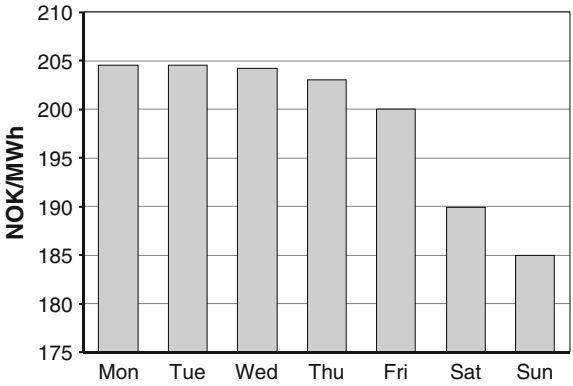


Fig. 3 Mean price (NOK/MWh) over the week, 1995–2007

significantly lower – and even more so on Saturday and Sunday. Thus, the Monday price has on the average topped the Sunday price by approximately 11%.

Table 1 summarizes the percentage changes (dLog) from 1 day of the week to the next for the entire period as well as subperiods. As can be seen, the price change from Sunday to Monday has been between 9 and 12%. For the entire sample (626 weeks), there has been a further significant price increase from Monday to Tuesday. This, however, seems to be a pattern that disappears after 2001 when there is no longer a significant price change from Monday till Tuesday. Friday, except for the period 2001–2003, shows significant negative price changes at the start of the

Table 1 Mean percentage price changes since the previous day

	2 Jan 1995–31 Dec 2006	2 Jan 1995–1 Jan 1998	1 Jan 2001–1 Jan 2004	1 Jan 2004–31 Dec 2006
Monday	0.113 (34.20)**	0.095 (14.30)**	0.121 (17.10)**	0.092 (20.40)**
Tuesday	0.008 (2.32)*	0.015 (2.27)*	0.001 (0.184)	−0.002 (−0.346)
Wednesday	−0.002 (−0.506)	−0.009 (−1.34)	−0.002 (−0.246)	0.004 (0.936)
Thursday	−0.007 (−2.24)*	−0.000 (−0.09)	−0.012 (−1.63)	−0.004 (−0.819)
Friday	−0.018 (−5.45)**	−0.014 (−2.06)*	−0.006 (−0.883)	−0.021 (−4.66)**
Saturday	−0.063 (−18.90)**	−0.056 (−8.39)**	−0.071 (−9.93)**	−0.040 (−8.86)**
Sunday	−0.031 (−9.31)**	−0.032 (−4.82)**	−0.029 (−4.11)**	−0.030 (−6.57)**
N	4,382 days; 625 weeks	1,096 days; 156 weeks	1,096 days; 156 weeks	1,096 days; 156 weeks

( ) = *t*-values
*/** Significant at 0.05/0.01 level

week-end, followed by substantial price reductions on Saturday and Sunday (−6.3% on Saturday and −3.1% on Sunday for the whole sample).

Despite some minor differences across the three subperiods, the week-day pattern appears to be very stable. On the margin, a producer who could move the supply of X MWh from Sunday to Monday would have a return of some 9–12% over the 24 h. A consumer who could move her demand from, say, Saturday to Sunday would save roughly 3% and as much as 9% if she could postpone her consumption from Friday till Sunday. There is reason to ask whether consumers and producers are taking full advantage of these strong short-term price regularities.

4 The Week-End Pattern

Within the day-to-day pattern, the price movement through the week-end reveals another interesting pattern. Power prices indicate that the week-end has become more oval during the last few years. While Saturday and Sunday a few years ago represented the true week-end, Friday has gradually moved into the leisure zone. The development is illustrated in Fig. 4, showing the probability of a price increase on a Friday ±2 S.E. estimated recursively by continuously increasing the observation window from 1995 to 1997 and onwards. Up till 1997–1998 this probability was not significantly different from 0.5, i.e., till 1997–1998 the price change from Thursday to Friday might just as well be positive as negative. Then, the probability

Fig. 4 Probability of a price increase on Fridays, recursive estimation, initialization 1995–97

for a positive price increase on Friday declines significantly towards roughly 0.35. In other words, Friday gradually seems to have been integrated into the week-end price reduction. This is due to a relative decline in demand on Fridays, i.e., a gradual change in consumer behavior ("the oval week-end"). The figure understates the decline in positive price changes on Fridays. For the period 2000–2006 (2,557 days), there was a price decline 72% of all Fridays, as compared to 57% for all Thursdays and 43% for all days (including Fridays).

Again, there is reason to ask whether this pattern could be utilized for profitable arbitrage or change in production plans, e.g., by moving production volumes from Fridays to Thursdays.

5 Time-of-the-Day Pattern

The power market is characterized by a very regular pattern throughout the day. The pattern is visualized in Fig. 5 and summarized in Table 2. Just before midnight, the price drops by some 3%. This decline continues for another 4 h, generating an average price reduction of some 10–12% between 10 p.m. and 4 a.m. Then, between 5 a.m. and 9 a.m. there is a significant price surge, averaging 15–18%. This pattern has remained quite stable since 1995, although with some peculiarities during the subperiod 2001–2004 when there were some ups and downs during the afternoon.

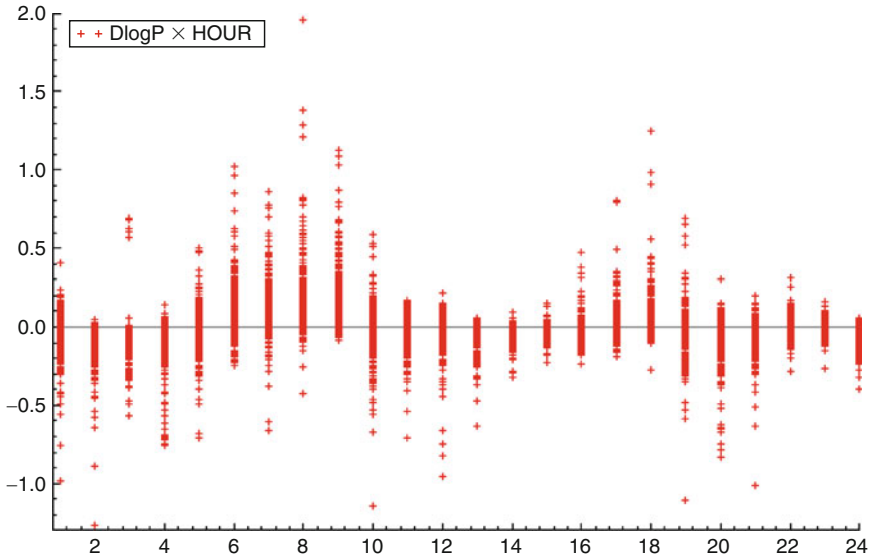


Fig. 5 Percentage price changes around the clock, 2000–2007; $N = 61{,}361$

Table 2 Mean percentage price change since the previous hour

Hour	1995–2006	1995–1998	2001–2004	2004–2006
1	−1.0	−1.7	−0.7	−0.6
2	**−2.7**	**−2.5**	**−2.9**	**−2.4**
3	−1.7	−1.1	−1.8	−1.9
4	−1.3	−0.8	−1.3	−1.6
5	0.5	0.1	0.4	0.5
6	**3.2**	**2.2**	**3.3**	**3.3**
7	**4.4**	**4.7**	**4.3**	**3.6**
8	**5.5**	**4.6**	**6.5**	**4.4**
9	**4.8**	**2.5**	**6.3**	**4.1**
10	−0.3	0.6	−1.0	−0.2
11	−0.2	0.0	−0.1	0.0
12	−0.9	−0.5	−1.1	−0.6
13	−1.6	−1.0	**−2.3**	−1.2
14	−1.1	−0.6	−1.3	−1.0
15	−0.8	−0.4	−1.0	−0.7
16	−0.3	−0.4	−0.2	−0.3
17	0.6	0.0	1.5	0.5
18	1.8	0.8	**2.5**	1.5
19	−0.2	0.0	−0.2	0.1
20	−1.5	−0.5	**−2.6**	−1.2
21	−1.4	−0.6	**−2.2**	−1.2
22	−0.4	−0.1	−0.4	−0.6
23	−1.5	−1.8	−1.4	−1.2
24	**−3.3**	**−3.2**	**−3.3**	**−3.0**

Bold types: Significant at 0.01 level

One must ask whether such extreme and regular short-term amplitudes are consistent with an efficient market. Would it not be possible to make profits from moving production or consumption by just a few hours?

6 Mean Reversion

Previous studies (e.g., Schwartz 1997; Weron et al. 2004; Cartea and Figueroa 2005; Simonsen et al. 2004) have found electricity prices to be mean reverting. Given the observed regularities over the day and during the week, this is no surprise. However, one would assume that the mean reverting pattern would disappear when adjusting for these calendar and time-of-the-day regularities, or when looking at time series for which there is no reason to find such cycles. This can be done by estimating models of the type

$$p_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i \, p_{t-i} + \varepsilon \tag{1}$$

$$p_t = \sum_{i=1}^{n} \alpha_i \, p_{t-i} + \sum_{i=1}^{7} \lambda_i \, DAY_i + \varepsilon_t \tag{2}$$

where p_t is the percentage price change since last week (1) or since yesterday (2) and where DAY is a (1;0)-dummy for each week day (Monday, Tuesday, etc.). For (1), we would not expect to find any systematic pattern in the way that the price change from 1 week to the next is related to previous weekly price changes. For (2), we would expect to find no significant alpha-estimates, having "taken out" the week-day regularity.

As can be seen from Tables 3 and 4, both hypotheses are rejected. There is a significant relationship between the percentage price change this week and the change that took place 1 and 2 weeks ago. True, the explained variance is very small (7% for all periods). Still, the regularity *may* be used at least for forecasting the *direction* of the price 1 week to the next.[3] As regards the daily price changes, a significant mean reversion still remains after correcting for the week day effect. Thus, there is a 7-day cycle with a significant and positive parameter estimates for the price change (t–7) and then significant and negative estimates for (t–1) through (t–5). The explained variance is small. Still, there is an additional explanation, although small, by including the previous 7 days' price changes in addition to the week-day dummies.

[3] We conducted a small experiment, using model (1) for weekly forecasting out-of-sample. During the 104 weeks 2005–2006, we predicted the correct direction during 63 weeks. With an a priori probability of 0.5, this result is relatively unlikely to occur by chance (0.02 probability).

Table 3 Estimation results model (1), weekly observations

	α_0	α_1	α_2	DW	R^2
1995 (4)–2006 (52)	0.0008 (0.183)	0.12 (3.09)	−0.24 (−6.24)	1.98	0.07
1995 (4)–1999 (52)	−0.0008 (−0.08)	0.13 (2.11)	−0.26 (−4.30)	2.00	0.08
2000 (1)–2004 (52)	0.002 (0.28)	0.10 (1.65)	−0.25 (−4.23)	1.94	0.07
2002 (1)–2006 (52)	0.001 (0.252)	0.23 (3.83)	−0.16 (−2.68)	1.89	0.07

( ) = t-values

Table 4 Estimation results, model (2), daily observations

	Estimated parameter 1995–2006	t-value	Estimated parameter 2003–2006	t-value
P^*_{t-1}	−0.13	−8.62**	−0.14	−5.40**
P^*_{t-2}	−0.14	−9.24**	−0.13	−4.91**
P^*_{t-3}	−0.07	−4.38**	−0.10	−3.66**
P^*_{t-4}	−0.09	−6.19**	−0.10	−3.99**
P^*_{t-5}	−0.08	−5.14**	−0.09	−3.22**
P^*_{t-6}	0.01	0.74**	0.01	0.84
P^*_{t-7}	0.13	8.39	0.13	3.50**
Constant	0.00	0.00	0.00	0.00
DW	2.01		2.01	
R^2	0.06		0.06	
N	4,375		1,461	

The estimated lambdas (week day means) are not reported
** Significant at 0.01 level

7 Tentative Conclusions

As stated in the introduction: it comes as no surprise to find time regularities in electricity prices. However, the strength, persistence, and magnitude of the regularities observed raise the question whether these regularities may offer profitable trading/production planning strategies or also profit opportunities from investments in technologies that allow for greater flexibility in energy production and/or consumption. Some investments aiming at utilizing intertemporal price differences are quite expensive, e.g., investments in pumping water back into the reservoir, investments in increased reservoir capacity or in new "complementary" production technologies. On the other hand, not so expensive technologies on the consumption side are available that may make it profitable to change consumption behavior, such as investments in dual burners (e.g., electricity/oil) and investments in metering technology for RTP. Whether the observed regularities are sufficiently strong for supporting such investments in new technologies remains to be seen. However, just by presenting detailed information on price regularities, one may see that the regularities become less pronounced.

Acknowledgment The author is grateful to participants at various seminars where earlier versions of this paper have been presented and to the anonymous referee for useful and helpful comments.

References

Bessembinder, H., & Lemmon, M. L. (2002). Equilibrium pricing and optimal hedging in electricity forward markets. *The Journal of Finance, 57*(3), 1347–1382

Borenstein, S. (2009). Electricity pricing that reflects its real-time cost. *NBER Reporter, 1*, 9–12

Burger, M. M., Klar, B., Müller, A., & Schindlmayr, G. (2004). A spot market model for pricing derivatives in electricity markets. *Quantitative Finance, 4*(1), 109–122

Byström, H. N. E. (2003). The hedging performance of electricity futures on the Nordic power exchange. *Applied Economics, 35*, 1–11

Cartea, A., & Villaplana, P. (2008). Spot price modeling and the valuation of electricity forward contracts: The role of demand and capacity. *Journal of Banking and Finance, 32*, 2502–2519

Cartea, A., & Figueroa, M. G. (2005). Pricing in electricity markets: A mean reverting jump diffusion model with seasonality. *Applied Mathematical finance, 12*, 313–335

Conejo, A. J., Contreras, J., Espinola, R., & Plazas, M. A. (2005). Forecasting electricity prices for a day-ahead pool-based electric energy market. *International Journal of Forecasting, 21*, 435–462

Gareta, R., Romeo, L. M., & Gil, A. (2005). Forecasting of electricity prices with neural networks. *Energy Conversion and Management, 47*, 1770–1778

Gjølberg, O., & Johnsen, T. (2001). Electricity futures: Inventories and price relationships at Nord Pool. Discussion Paper IOR/UMB #D-16/2001

Koekebakker, S., & Ollmar, F. (2005). Forward curve dynamics in the Nordic electricity market. *Managerial Finance, 31*(6), 74–95

Kristiansen, T. (2007). Pricing of monthly forward contracts in the Nord Pool market. *Energy Policy, 35*, 307–316

Longstaff, F. A., & Wang, A. W. (2004). Electricity forward prices: A high-frequency empirical analysis. *Journal of finance, LIX*(4), 1887–1900

Lucia, J. J., & Schwartz, E. S. (2002). Electricity prices and power derivatives: Evidence from the Nordic power exchange. *Review of Derivative Research, 5*, 5–50

Park, H., Mjelde, J. W., & Bessler, D. A. (2006). Price dynamics among U.S. electricity spot markets. *Energy Economics, 28*, 81–101

Robinson, T., & Baniak, A. (2002). The volatility of prices in the English and Welsh electricity pool. *Applied Economics, 34*, 1487–1495

Shawky, H. A., Marathe, A., & Barrett, C. L. (2003). A first look at the empirical relation between spot and futures electricity prices in the United States. *Journal of Futures markets, 23*(10), 931–955

Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance, 52*, 923–974

Simonsen, I., Weron, R., Mo, B. (2004). Structure and stylized facts of a deregulated power market. *MPRA Paper* No. 1443, http://mpra.ub.uni-muenchen.de/1443/

Simonsen, I. (2005). Volatility of power markets. *Physica A*, *355*, 10–20

Weron, R. (2000). Energy price risk management. *Physica A, 285*, 127–134

Weron, R., Simonsen, I., & Wilman, P. (2004). Modeling highly volatile and seasonal markets: Evidence from the Nord Pool electricity market. In H. Takayasu (Ed.), *Toward control of economic change – Applications of econophysics* (pp. 182–191). Tokyo: Springer-Verlag

Wilkens, S., & Wimschulte, J. (2007). The pricing of electricity futures: Evidence from the European energy exchange. *Journal of Futures markets, 27*(4), 387–410

Woo, C.- K., Horowitz, I., & Hoang, K. (2001). Cross hedging and forward-contract pricing of electricity. *Energy Economics, 23*, 1–15

Worthington, A., Kay-Spratley, A., & Higgs, H. (2005). Transmission of prices and price volatility in Australian electricity spot markets: A multivariate GARCH analysis. *Energy Economics, 27*, 337–350

Yamin, H. Y., Shahidehpour, S. M., & Li, Z. (2004). Adaptive short-term electricity price forecasting using artificial neural networks in the restructured power markets. *Journal of Electrical Power and Energy Systems, 8*, 571–581

Valuation and Risk Management in the Norwegian Electricity Market*

Petter Bjerksund, Heine Rasmussen, and Gunnar Stensland

Abstract The purpose of this paper is twofold: Firstly, we analyse the option value approximation of traded options in the presence of a volatility term structure. The options are identified as: (a) "European" (written on the forward price of a future flow delivery); and (b) "Asian". Both types are in fact written on (arithmetic) price averages. Secondly, adopting a 3-factor model for market risk which is compatible with the valuation results, we discuss risk management in the electricity market within the Value at Risk concept. The analysis is illustrated by numerical cases from the Norwegian electricity derivatives market.

1 Introduction

Historical time series, implicit volatilities of quoted option prices, as well as the experience of professional traders and brokers, clearly indicate the presence of a volatility term structure in the Norwegian electricity derivatives market. The purpose of this paper is to analyse the implications of this volatility term structure for: (a) valuation of the most frequently traded options; and (b) market risk management.

Our starting point is to represent the electricity forward market at date t by a forward price function $f(t, T)$, which may be interpreted as the forward price at date t of a hypothetical contract with delivery at date T (i.e., with an infinitesemal delivery period). In the electricity forward market, the underlying quantity is delivered as a flow during a specific future time period. This contract may be interpreted as a portfolio of hypothetical single-delivery contracts, hence the forward price follows from the function $f(t, T)$ by no-arbitrage.

*This chapter is a corrected version of Bjerksund, Rasmussen, and Stensland (2000).

P. Bjerksund (✉) and G. Stensland
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: petter.bjerksund@nhh.no; gunnar.stensland@nhh.no

H. Rasmussen
Statkraft, 0216 Oslo, Norway
e-mail: heine.rasmussen@statkraft.com

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_11,
© Springer-Verlag Berlin Heidelberg 2010

Assuming lognormality, we represent the uncertainty in the forward market at date t by a volatility function $\sigma(\tau - t, T - t)$, which corresponds to the Black'76 implicit volatility of a European option with time to exercise $\tau - t$ written on the future forward price $f(t, T)$ with time to delivery $T - t$.

However, the traded "European" electricity option is written on the forward price of a contract with delivery as a constant flow during a specific future time period. Following Kemna and Vorst (1990), we adopt the Black'76 concept for approximating the option value and obtain the theoretical forward price as well as an approximated plug-in volatility.

The traded Asian option is written on the average spot price observed during a specific period. The exercise date of the option typically coincides with the last observation date. We obtain the theoretical forward price and the Black'76 plug-in volatility.

Next, we turn to risk management within the Value at Risk concept. The idea of Value at Risk is to quantify the downside risk of the future market value of a given portfolio at a chosen horizon date. We represent the market risk by a 3-factor model, which is compatible with our forward price dynamics assumption. We use Monte Carlo simulation in order to generate the probability distribution of the future portfolio market price.

The advantage of integrating valuation and risk management is: (a) the market risk exposure of a future position is consistent with the current forward and option prices; and (b) we may use our option valuation approximation results to calculate conditional future option values.

2 The Model

2.1 The Forward Market

Research on valuation of commodity derivatives and management of commodity market risk has been an expanding area within finance during the last decade. At the same time, the use of various bilateral OTC arrangements in the industry has increased, and new commodity derivatives have been introduced in the financial market place.

For many commodities, the forward prices indicate a non-constant convenience yield (e.g., seasonal pattern). Moreover, the commodity option market prices clearly indicate that the constant volatility assumption of Black'76 is violated for most commodities. Typically, the implicit volatility is a decreasing and convex function of time to maturity.

Gibson and Schwartz (1990) develop a two-factor model for oil derivatives, where the commodity spot price is geometric Brownian, and the instantaneous convenience yield rate follows a mean-reverting Ornstein–Uhlenbeck process. Within this model, closed form solutions exist for the forward price as well as European

calls (see Bjerksund (1991) and Jamshidian and Fein (1990)). Hilliard and Reis (1998) investigate several alternative models, including the case where the spot price is a mixed jump-diffusion process. For a survey on alternative models for valuation and hedging, see Schwartz (1997).

Models where assumptions on spot price and convenience yield dynamics are starting points will typically predict forward prices which are different from the ones observed in the market. Using the general Heath, Jarrow, and Morton (1992) approach, Miltersen and Schwartz (1998) develop a general framework for commodity derivatives valuation and risk management with stochastic interest rates as well as stochastic convenience yield. This model can be calibrated to the current forward market. In their Gaussian special case, the call option value essentially boils down to a generalised version of Black'76.

Our model assumptions may be considered as a special case of the gaussian Miltersen–Schwartz model. Complicating the picture in the case of electricity derivatives, however, is the fact that the physical "underlying asset" is a constant flow received during a specific time period, rather than one "bulk" delivery at a specific date.

Turning to our model, we represent the forward market at date t by a continuous forward price function, where $f(t,T)$ denotes the forward price at date t on a contract with delivery at date $T \geq t$. Consider a forward contract with delivery date T, and assume the following forward price dynamics at date $t \leq T$ (with respect to the risk-adjusted martingale probability measure)

$$\frac{df(t,T)}{f(t,T)} = \left(\frac{a}{T-t+b} + c\right) dW^*(t), \tag{1}$$

where a, b, and c are positive constants, and $dW^*(t)$ is the increment of a standard Brownian motion with expectation $E_t^*[dW^*(t)] = 0$ and $\mathrm{Var}_t^*[dW^*(t)] = dt$. By construction, the expectation of (1) is zero with respect to the martingale measure.

The above corresponds to the forward price of this contract at the future date $\tau \in [t,T]$ being lognormal, and given by the following stochastic integral

$$f(\tau,T) = f(t,T) \exp\left\{\int_t^\tau \left(\frac{a}{T-s+b} + c\right) dW^*(s) - \tfrac{1}{2}\int_t^\tau \left(\frac{a}{T-s+b} + c\right)^2 ds\right\}.$$

Observe that $E_t^*[1_\tau f(\tau,T)] = f(t,T)$, which confirms that the forward price is a martingale with respect to the $*$-probability measure.

Now, consider a hypothetical European call option with time to exercise $\tau - t$, written on the future forward price $f(\tau,T)$ on a contract with time to delivery $T-t$. It follows from the literature (see, e.g., Harrison and Kreps (1979) and Harrison and Pliska (1981)) that the market value of the option can be represented by the expected (using the martingale measure) discounted (using the riskless rate) future

pay-off. With the future forward price being lognormal, the call value is given by the Black'76 formula

$$V_t\left[1_\tau\left(f(\tau,T)-K\right)^+\right]=E_t^*\left[e^{-r(\tau-t)}\left(f(\tau,T)-K\right)^+\right]$$
$$=e^{-r(\tau-t)}\left\{f(t,T)N(d_1)-KN(d_2)\right\}, \tag{2}$$

where $N(\cdot)$ is the standard normal cumulative probability function,

$$d_1\equiv\frac{\ln(f(t,T)/K)+\frac{1}{2}\sigma^2(\tau-t)}{\sigma\sqrt{\tau-t}}, \tag{3}$$

$$d_2\equiv d_1-\sigma\sqrt{\tau-t}, \tag{4}$$

$$\sigma\equiv\sqrt{\mathrm{Var}_t^*\left[\ln\left(\frac{f(\tau,T)}{f(t,T)}\right)\right]\Big/(\tau-t)}. \tag{5}$$

Observe that the key input of Black'76 is: (a) the forward price at date t of the underlying asset $f(t,T)$; and (b) the uncertainty of the underlying asset, represented by the volatility σ.

The assumed dynamics translates into the volatility σ being a function of time to exercise (of the option), $\tau-t$, and time to delivery (of the underlying forward), $T-t$, and given by

$$\sigma=\sigma(\tau-t,T-t)$$
$$=\sqrt{\mathrm{Var}_t^*\left[\ln\left(\frac{f(\tau,T)}{f(t,T)}\right)\right]\Big/(\tau-t)}, \tag{6}$$

where[1]

$$\mathrm{Var}_t^*\left[\ln\left(\frac{f(\tau,T)}{f(t,T)}\right)\right]=\mathrm{Var}_t^*\left[\int_{s=t}^{s=\tau}\frac{df(s,T)}{f(s,T)}\right]$$
$$=\left[\frac{a^2}{T-s+b}-2ac\ln(T-s+b)+c^2s\right]_{s=t}^{s=\tau}. \tag{7}$$

In the following, we represent the forward market at date t by the forward price function $f(t,T)$ and the volatility function $\sigma(\tau-t,T-t)$.

[1] To establish the first equality, apply Ito's lemma

$$\mathrm{Var}_t^*\left[\ln\left(\frac{f(\tau,T)}{f(t,T)}\right)\right]=\mathrm{Var}_t^*\left[\int_{s=t}^{s=\tau}\left(\frac{df(s,T)}{f(s,T)}\right)-\int_{s=t}^{s=\tau}\frac{1}{2}\left(\frac{df(s,T)}{f(s,T)}\right)^2\right],$$

insert the assumed forward price dynamics, and observe that the second integral is deterministic as of date t. The second equality follows from the fact that Brownian motions have independent increments across time.

3 European Option

3.1 Forward on a Flow Delivery

In the electricity forward market, the underlying physical commodity is delivered during a specific time period $[T_1, T_2]$ as a constant flow (at a rate of $(T_2 - T_1)^{-1}$ units per year). We observe delivery periods on contracts ranging from one day to one year, depending on the remaining time to delivery of the contract.

We represent the forward market at date t by the forward price function $f(t, s)$, $t \leq s \leq T$. By value additivity, the market value at date t of receiving one unit of the commodity from dates T_1 to T_2 (at a rate of $1/(T_2 - T_1)$) is simply

$$V_t \left[\int_{T_1}^{T_2} 1_s \frac{f(s,s)}{T_2 - T_1} ds \right] = \int_{T_1}^{T_2} e^{-r(s-t)} \frac{f(t,s)}{T_2 - T_1} ds, \tag{8}$$

where $t \leq T_1 < T_2$. In a rational market, the forward price $F(t, T_1, T_2)$ is determined such that the market value at date t of the payments equals the right-hand side of the equation just above. Indeed, in the hypothetical case of up-front payment at date t, the forward price would coincide with the right-hand side just above.

Now, suppose that the forward price is paid as a constant cash flow stream during the delivery period (at a rate of $F(t, T_1, T_2)/(T_2 - T_1)$ per time unit). At date t, the net market value of entering the contract is zero, leading to the following forward price

$$F(t, T_1, T_2) = \int_{T_1}^{T_2} w(s; r) f(t, s) ds, \tag{9}$$

where

$$w(s; r) = \frac{e^{-rs}}{\int_{T_1}^{T_2} e^{-rs} ds}. \tag{10}$$

Consequently, the forward price $F(t, T_1, T_2)$ may be interpreted as the average of the forward prices $f(t, s)$ over the delivery period $[T_1, T_2]$, with respect to the weight function[2], which reflects the time value of money.

3.2 Call Option Valuation

The European calls which are traded in the electricity derivatives market are typically written on a forward price. In particular, consider a European call option written on the pay-off $F(\tau, T_1, T_2)$ with strike K and exercise date $\tau \leq T_1$. Observe that the exercise date of the option precedes the delivery period of the underlying forward contract.

[2] Observe that $w(s; r) > 0 \forall s \in [T_1, T_2]$ and $\int_{T_1}^{T_2} w(s; r) ds = 1$.

Following Kemna and Vorst (op.cit.), we approximate the option value within the Black'76 framework. We have already obtained the theoretical forward price of the underlying uncertain pay-off, $F(t, T_1, T_2)$. In addition, we need an approximated volatility parameter. Approximate the forward price dynamics for $t \le T_1$ by[3]

$$\begin{aligned}\frac{dF(t,T_1,T_2)}{F(t,T_1,T_2)} &\approx \int_{s=T_1}^{s=T_2} \frac{1}{T_2-T_1}\frac{df(t,s)}{f(t,s)}ds \\ &= \left\{\frac{a}{T_2-T_1}\ln\left(\frac{T_2-t+b}{T_1-t+b}\right)+c\right\}dW^*(t). \end{aligned} \tag{11}$$

Next, obtain the approximated variance

$$\begin{aligned}\mathrm{Var}_t^*\left[\ln\left(\frac{F(\tau,T_1,T_2)}{F(t,T_1,T_2}\right)\right] &= \mathrm{Var}_t^*\left[\int_t^\tau \frac{dF(s,T_1,T_2)}{F(s,T_1,T_2)}ds\right] \\ &= \left(\frac{a}{T_2-T_1}\right)^2\int_t^\tau\left(\ln\frac{T_2-s+b}{T_1-s+b}\right)^2 ds \\ &+ \frac{2ac}{T_2-T_1}\int_t^\tau \ln\frac{T_2-s+b}{T_1-s+b}ds + c^2\int_t^\tau ds, \end{aligned} \tag{12}$$

where the first and the second integrals are[4]

$$\begin{aligned}\int_t^\tau\left(\ln\frac{T_2-s+b}{T_1-s+b}\right)^2 ds = \Big[&(x+\alpha)\,(\ln(x+\alpha))^2 \\ &-2(x+\alpha)\ln(x+\alpha)\ln(x-\alpha) \\ &+4\alpha\ln(2\alpha)\ln\left(\frac{x-\alpha}{2\alpha}\right) \\ &-4\alpha\mathrm{dilog}\left(\frac{x+\alpha}{2\alpha}\right) \\ &+(x-\alpha)\,(\ln(x-\alpha))^2\Big]_{X(\tau)}^{X(t)}, \end{aligned} \tag{13}$$

$$\begin{aligned}\int_t^\tau \ln\frac{T_2-s+b}{T_1-s+b}ds = \big[&(x+\alpha)\ln(x+\alpha) \\ &-(x-\alpha)\ln(x-\alpha)\big]_{X(\tau)}^{X(t)}, \end{aligned} \tag{14}$$

[3] The approximation proceeds in the following two steps

$$\frac{dF(t,T_1,T_2)}{F(t,T_1,T_2)} \approx \int_{s=T_1}^{s=T_2} w(s;r)\frac{df(t,s)}{f(t,s)}ds \approx \int_{s=T_1}^{s=T_2} w(s;0)\frac{df(t,s)}{f(t,s)}ds.$$

[4] We have corrected the typos in a previous version of this paper pointed out in Lindell and Raab (2008).

where we define

$$\alpha \equiv \tfrac{1}{2}(T_2 - T_1), \tag{15}$$

$$X(s) \equiv b + \tfrac{1}{2}(T_2 + T_1) - s, \tag{16}$$

and where the dilogarithm function is defined by[5]

$$\text{dilog}\,(x) = \int_1^x \frac{\ln(s)}{1-s} ds \text{ where } x \geq 0 \tag{17}$$

see, for example, Abramowitz and Stegun (1972).

Now, consider a European call option with exercise date τ written on the forward price $F(\tau, T_1, T_2)$, where $t < \tau \leq T_1 < T_2$ The option value at date t can now be approximated by Black'76, using the forward price $F(t, T_1, T_2)$ above and the volatility parameter v_E

$$\begin{aligned} v_E &\equiv v_E(\tau - t, T_1 - t, T_2 - t) \\ &= \sqrt{\text{Var}_t^* \left[\ln\left(\frac{F(\tau, T_1, T_2)}{F(t, T_1, T_2)} \right) \right] \Big/ (\tau - t)} \end{aligned} \tag{18}$$

The volatility parameter v_E associated with the European option is a function of the time to maturity of the option $(\tau - t)$, the time to start of delivery $(T_1 - t)$, and the time to stop of delivery $(T_2 - t)$.

4 Asian Option

Asian options are written on the *average* spot price observed during a specific period $[T_1, T_2]$, with exercise date $\tau \geq T_2$. With continuous sampling, the (arithmetic) average of the spot prices $f(s, s)$ observed from T_1 to T_2 is defined by

$$A(T_1, T_2) \equiv \int_{T_1}^{T_2} \frac{1}{T_2 - T_1} f(s, s) ds. \tag{19}$$

We are interested in evaluating a call option with strike K and exercise date T_2, written on the arithmetic average $A(T_1, T_2)$. For simplicity, we deal with the case of

[5] The function is approximated numerically by

$$\text{dilog}\,(x) = \begin{cases} \sum_{k=1}^{n} \frac{(x-1)^k}{k^2} & \text{for } 0 \leq x \leq 1 \\ -\frac{1}{2}(\ln(x))^2 - \sum_{k=1}^{n} \frac{((1/x)-1)^k}{k^2} & \text{for } x > 1 \end{cases}$$

where n is a sufficiently large positive integer.

$t \leq T_1$ first. With the future spot prices being lognormal, there is no known probability distribution for the arithmetic average. Within the Black'76 framework, the option value approximation problem boils down to finding the theoretical forward price and a reasonable volatility parameter.

Now, it follows from the martingale property of forward prices that the forward price on a contract written on (the cash equivalent of) $A(T_1, T_2)$ with delivery at date T_2 is

$$\begin{aligned} F_t[A(T_1,T_2)] &= E_t^* \left[\int_{T_1}^{T_2} \frac{1}{T_2 - T_1} f(s,s) ds \right] \\ &= \int_{T_1}^{T_2} \frac{1}{T_2 - T_1} f(t,s) ds. \end{aligned} \tag{20}$$

Observe that the forward price $F_t[A(T_1, T_2)]$ simply is the (equally weighted) arithmetic average of the current forward prices over the sampling period $[T_1, T_2]$. This forward price may be interpreted as the cost replicating this contract in the market.[6]

Turning to the Black'76 volatility parameter, approximate the dynamics of the underlying forward price at date $\tau \in [t, T_2]$ by

$$\begin{aligned} \frac{dF_\tau[A(T_1,T_2)]}{F_\tau[A(T_1,T_2)]} &\approx \int_{s=\max\{t,T_1\}}^{s=T_2} \frac{1}{T_2 - T_1} \frac{df(\tau,s)}{f(\tau,s)} ds \\ &= \begin{cases} \left\{ \dfrac{a}{T_2 - T_1} \ln\left(\dfrac{T_2 - \tau + b}{T_1 - \tau + b} \right) + c \right\} dW^*(\tau) & \text{when } \tau \leq T_1 \\ \left\{ \dfrac{a}{T_2 - T_1} \ln\left(\dfrac{T_2 - \tau + b}{b} \right) + \dfrac{T_2 - \tau}{T_2 - T_1} c \right\} dW^*(\tau) & \text{when } \tau > T_1 \end{cases} \end{aligned} \tag{21}$$

Obtain the approximated variance by

$$\begin{aligned} &\mathrm{Var}_t^* \left[\ln\left(\frac{A(T_1,T_2)}{F_t[A(T_1,T_2)]} \right) \right] \\ &= \mathrm{Var}_t^* \left[\int_{\tau=t}^{\tau=T_2} \frac{dF_\tau[A(T_1,T_2)]}{F_\tau[A(T_1,T_2)]} ds \right] \\ &= \left(\frac{a}{T_2 - T_1} \right)^2 \int_t^{T_1} \left(\ln \frac{T_2 - \tau + b}{T_1 - \tau + b} \right)^2 d\tau \end{aligned}$$

[6] Assume for the moment a discrete time model where the delivery period $[T_1, T_2]$ is divided into n time intervals of time length Δt. Consider the following strategy: At the evaluation date t, buy $e^{-r(T_2-(T_1+i\cdot\Delta t))}(1/n)$ units forward for each delivery $T_1 + i \cdot \Delta t$, $i = 1, \ldots, n$. As time passes and the contracts are settled, invest (or finance) the proceeds at the riskless interest rate r. At the delivery date $\tau \geq T_2$, the pay-off from the strategy is $\sum_{i=1}^n (1/n) f(T_1 + i \cdot \Delta t, T_1 + i \cdot \Delta t) - \sum_{i=1}^n (1/n) f(t, T_1 + i \cdot \Delta t)$, where the first term represents the desired spot price, and the second (riskless) term may be interpreted as the forward price as of date t.

$$+\frac{2ac}{T_2-T_1}\int_t^{T_1}\ln\frac{T_2-\tau+b}{T_1-\tau+b}d\tau+c^2\int_t^{T_1}d\tau \tag{22}$$
$$+\left(\frac{a}{T_2-T_1}\right)^2\int_{T_1}^{T_2}\left(\ln\frac{T_2-\tau+b}{b}\right)^2 d\tau$$
$$+\frac{2ac}{T_2-T_1}\int_{T_1}^{T_2}\ln\frac{T_2-\tau+b}{b}\frac{T_2-\tau}{T_2-T_1}d\tau+c^2\int_{T_1}^{T_2}\left(\frac{T_2-\tau}{T_2-T_1}\right)^2 d\tau,$$

where the first and the second integrals are evaluated by inserting $\tau = T_1$ in (13)–(14) above, and the fourth and the fifth integrals are

$$\int_{T_1}^{T_2}\left(\ln\frac{T_2-\tau+b}{b}\right)^2 d\tau = b\left[y\,(ln(y))^2-2y\ln(y)+2y\right]_1^{\overline{y}} \tag{23}$$

$$\int_{T_1}^{T_2}\ln\left(\frac{T_2-\tau+b}{b}\right)\frac{T_2-\tau}{T_2-T_1}d\tau=\frac{b^2\left[\frac{1}{2}y^2\ln(y)-y\ln(y)+y-\frac{1}{4}y^2\right]_1^{\overline{y}}}{T_2-T_1} \tag{24}$$

where

$$\overline{y}=\frac{T_2-T_1+b}{b}. \tag{25}$$

The Black'76 volatility parameter v_A is now found by

$$v_A \equiv v_A(T_1-t,T_2-t)$$
$$=\sqrt{\mathrm{Var}_t^*\left[\ln\left(\frac{A(T_1,T_2)}{F_t[A(T_1,T_2)]}\right)\right]\Big/(T_2-t)}. \tag{26}$$

Observe that the volatility parameter v_A is a function of time to the first sampling date, T_1-t, and time to the last sampling date, T_2-t, where the latter coincides with the time to exercise of the option.

Next, consider the case where the option is evaluated within the sampling period, that is, $T_1 < t \le T_2$. It follows immediately from the definition of the arithmetic average that

$$A(T_1,T_2)=\frac{t-T_1}{T_2-T_1}A(T_1,t)+\frac{T_2-t}{T_2-T_1}A(t,T_2). \tag{27}$$

Consequently, with $T_1 < t \le T_2$, the call option problem is equivalent to

$$V_t\left[1_{T_2}\left(A(T_1,T_2)-K\right)^+\right]=\frac{T_2-t}{T_2-T_1}V_t\left[1_{T_2}\left(A(t,T_2)-K'\right)^+\right], \tag{28}$$

where

$$K' \equiv \frac{T_2 - T_1}{T_2 - t} K - \frac{t - T_1}{T_2 - t} A(T_1, t), \tag{29}$$

that is, a portfolio of $\frac{T_2 - t}{T_2 - T_1}$ call options, each written on the average over the remaining sampling period $[t, T_2]$ where the strike is adjusted for the already observed prices. In the non-trivial case of $K' > 0$, the value of the adjusted option can be evaluated by inserting $T_1 = t$ and $K = K'$ in the evaluation procedure above. In the degenerate case of $K' \leq 0$, it will always be optimal to exercise the call, which reduces the adjusted option to a forward with the current value

$$V_t \left[1_{T_2} \left(A(t, T_2) - K' \right)^+ \right] = e^{-r(T_2 - t)} \left((T_2 - t)^{-1} \int_t^{T_2} f(t, s) ds - K' \right). \tag{30}$$

5 Valuation: An Example

5.1 *Current Term Structure*

The Nordic electricity market NORDPOOL consists of several forward and futures contracts. The traded contract and their market prices at December 15, 1999 are found in Fig. 1.

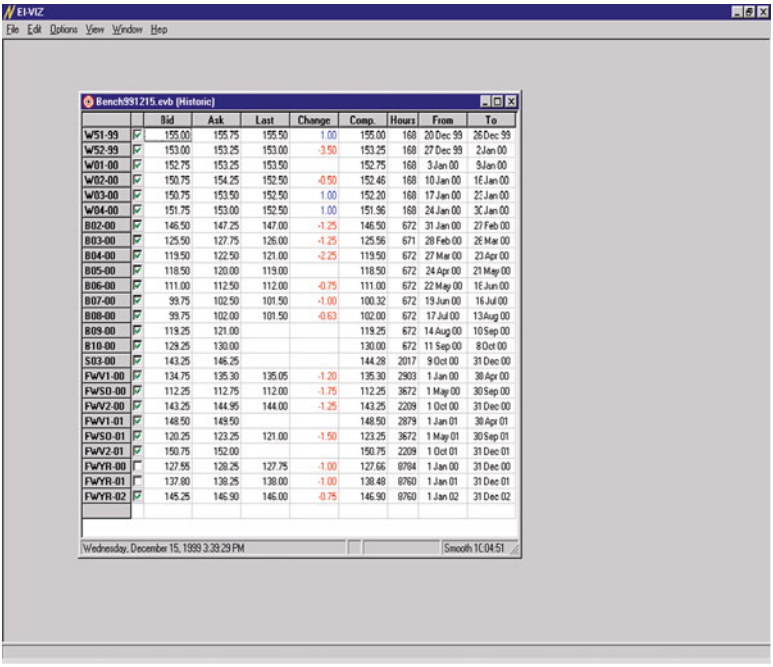


Fig. 1 Market prices at December 15, 1999

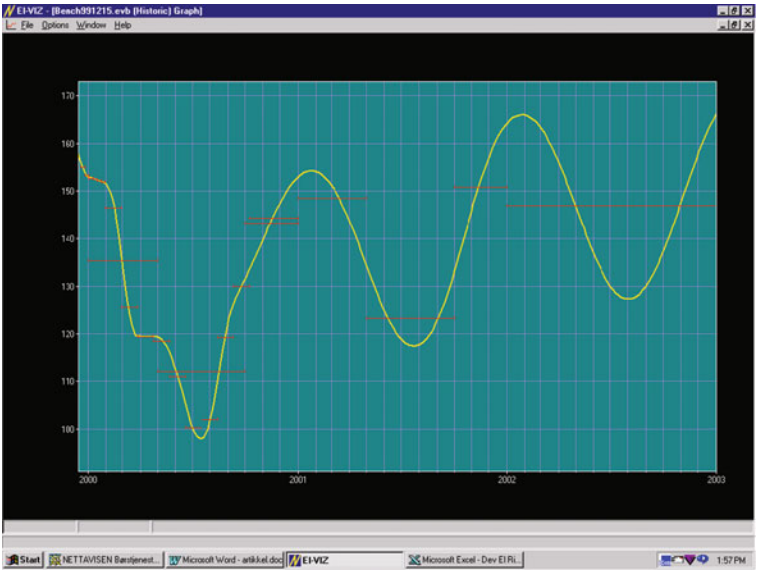


Fig. 2 Forward prices

On the Basis of the bid/ask prices, we construct a continuous forward price function. The forward function is given by the smoothest function that prices all traded contracts on NORDPOOL within the bid/ask spread. The forward price function on December 15, 1999 is represented by the continuous yellow curve in Fig. 2. The red horizontal lines in Fig. 2 correspond to the quoted forward price of each traded contract.

5.2 Volatility

The volatility in forward prices falls rapidly in this market. The volatility on a single day delivery starting in one week might be 80%, whereas a similar delivery starting in 6 months will typically have less than 20% immediate volatility.

Figure 3 shows the forward price function and the volatility curve at December 15, 1999 for the following calendar year (i.e., 2000).

5.3 Contract Valuation

In the following, we consider three valuation cases as of December 15, 1999. The first case corresponds to the contract "FWYR-2000 Asian/M", see the first line in Fig. 4. The strike of the option is 120 and the contract expires at December 31, 2000.

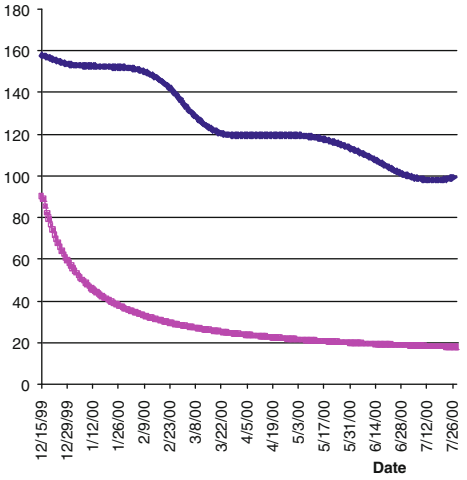


Fig. 3 Forward price and volatility

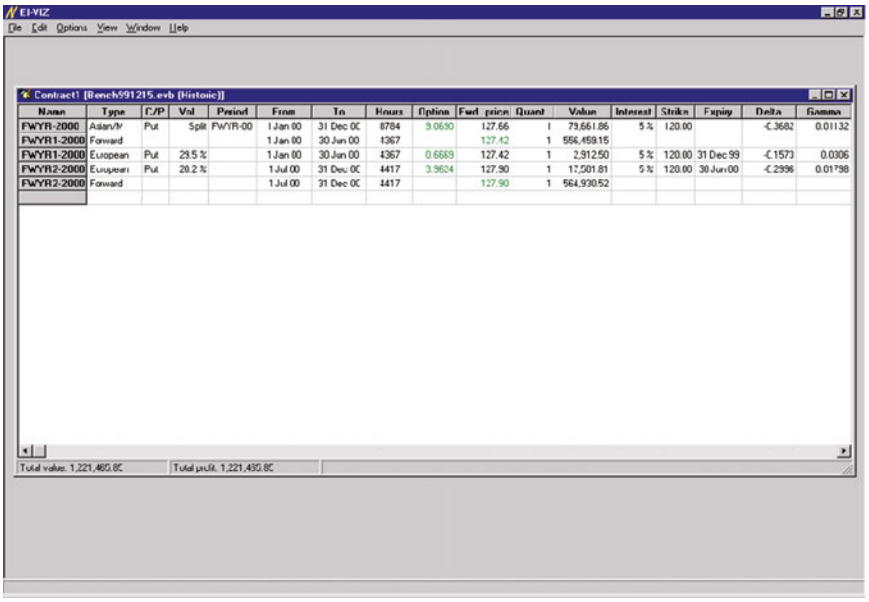


Fig. 4 Contract valuation

The contract is subject to "monthly settlements", which means that the contract represents a portfolio 12 monthly Asian options, where each option is written on the monthly price average and settled at the end of the month.

The second case is a European put option with strike 120 and expiration date December 31, 1999, written on the forward price on the forward contract on delivery from January 1, 2000 to June 30, 2000. The value of the option and the underlying contract are found in lines 3 and 2 in Fig. 4.

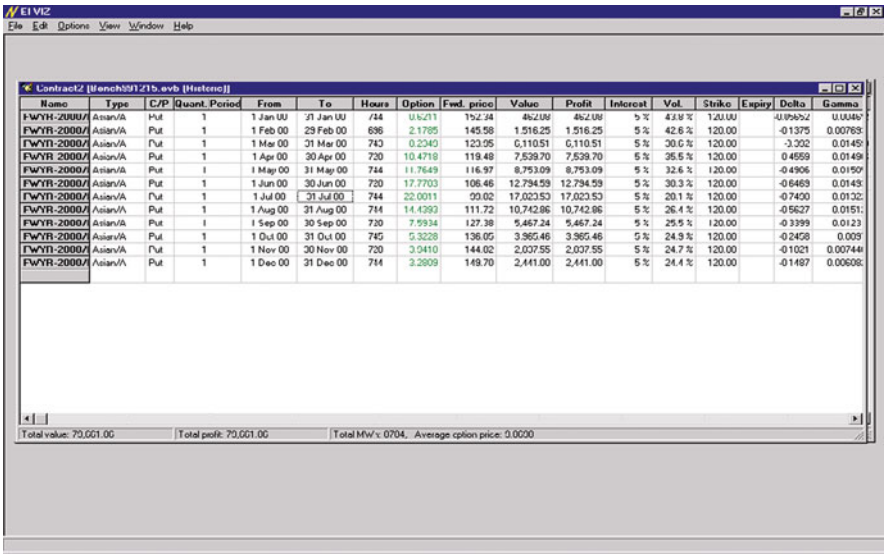


Fig. 5 Split of Asian option

The third case is a European put option with strike 120 and expiration date June 30, 2000, written on the forward price on the forward contract on delivery from July 1, 2000 to December 31, 2000. The value of the option and the underlying contract are found in lines 4 and 5.

Figure 5 considers the first case in more detail. Each line corresponds to an Asian option with strike 120 written on a monthly price average with expiration at the end of the month. Observe that as seen from December 15, 1999, the volatility of the underlying monthly price average is a decreasing and convex function of the delivery month (e.g., January 43.8%; June 30.3%; December 24.4%). By value additivity, the value of each monthly option adds up to the value of the quoted contract (79,661.86 in Fig. 4).

6 Value at Risk

The idea of Value at Risk (VaR) is to focus on the downside market risk of a given portfolio at a future horizon date. For a discussion on VaR, see Hull (1998) and Jorion (1997).

Evidence suggests that even though a one-factor model may be adequate for valuation in a multi-factor environment, it typically performs poorly as a tool for risk management (e.g., dynamic hedging). In the following, we discuss a three-factor Value at Risk (VaR) model, which is consistent with the valuation and approximation results above, following from (1) above.

In order to obtain a richer class of possible forward price functions, assume the following forward price dynamics (with respect to the martingale probability measure)

$$\frac{df(t,T)}{f(t,T)} = \frac{a}{T-t+b}dW_1^*(t) + \left(\frac{2ac}{T-t+b}\right)^{\frac{1}{2}} dW_2^*(t) + cdW_3^*(t), \qquad (31)$$

where a, b, and c are the positive constants from (1) above, and $dW_1^*(t)$, $dW_2^*(t)$, and $dW_3^*(t)$ are increments of three uncorrelated standard Brownian motions. Observe that the instantaneous dynamics of (31) just above is normal with zero expectation and variance

$$\mathrm{Var}_t^*\left[\frac{df(t,T)}{f(t,T)}\right] = \left\{\left(\frac{a}{T-t+b}\right)^2 + \frac{2ac}{T-t+b} + c^2\right\} ds, \qquad (32)$$

which is consistent with the dynamics of (1) above.

It follows that the forward price function $f(\tau,T)$ at the future date τ is the stochastic integral

$$\begin{aligned} f(\tau,T) = f(t,T)\exp&\left\{\int_t^\tau \frac{a}{T-s+b}dW_1^*(s) - \tfrac{1}{2}\int_t^\tau \left(\frac{a}{T-s+b}\right)^2 ds\right\} \\ &\exp\left\{\int_t^\tau \left(\frac{2ac}{T-s+b}\right)^{\frac{1}{2}} dW_2^*(s) - \tfrac{1}{2}\int_t^\tau \frac{2ac}{T-s+b}ds\right\} \qquad (33) \\ &\exp\left\{\int_t^\tau c \quad dW_3^*(s) - \tfrac{1}{2}\int_t^\tau c^2 ds\right\}. \end{aligned}$$

In addition, the forward market at the future date τ is represented by the associated Black'76 implicit volatility function $\sigma(\theta-\tau, T-\tau)$, where $\theta \in [\tau, T]$ is the exercise date of the option, and $T \geq \theta$ is the delivery date of the underlying forward.

Consider a portfolio of electricity derivatives at the future date τ. The idea of VaR is to analyse the downside properties of the probability distribution of the future portfolio value. We apply the simulation methodology in order to generate this probability distribution, from which Value at Risk can be calculated. The procedure consists of the following steps (which are repeated): First, use a random generator to draw a possible realisation for the future forward price function consistent with (33) above. Second, use the above valuation and approximation results to calculate the associated market value of each position, conditional on the realised forward price function (as well as the future implicit Black'76 volatility function). Thirdly, calculate the conditional market value of the portfolio (which follows immediately from value additivity). Now, for a large number of iterations, we approximate the probability distribution of the future portfolio value by the histogram following from the simulation results.

7 Value at Risk: An Example

7.1 Price Path Simulations

Equation (33) describes how the future forward price function is simulated from current market information. The $f(t, T)$ function is the forward price at time t for delivery at time T. The parameters a, b, and c are inputs to the volatility function.

In order to simulate possible price paths, we use (33) repeatedly. In Fig. 6 we present 100 simulated week prices based on this model. In each simulated path, the following procedure is followed. First, the forward function next week is simulated, integrating this curve from zero to 7 days gives the first week price. Next, we use this new forward curve in combination with the volatility curve to obtain the forward curve in the next step and so on. In this way, we obtain the correct and large short-term volatility in prices in addition to the much smaller volatility in prices as seen from today. We observe that the simulation model gives a substantial mean reversion in prices. This is in accordance with empirical data. The advantage of this method is that current information about the volatility curve and the term structure of prices is sufficient to perform this simulation.

7.2 Value at Risk Calculation

In the following, we focus on the downside risk of a given financial portfolio of forwards and options. Assume that we want a probability distribution which represents

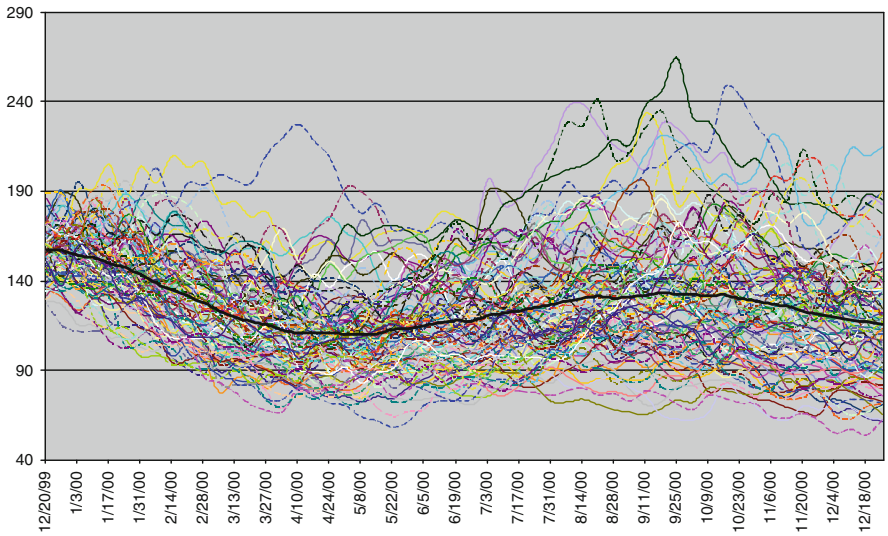


Fig. 6 Price path simulation

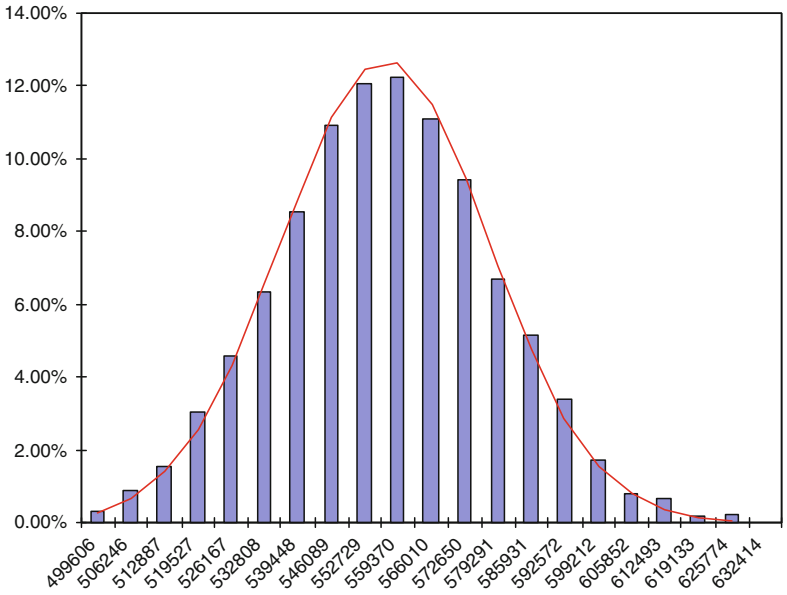


Fig. 7 Distribution for the value of the forward contract first half of 2000 in one week, NOK

the possible future market values of the portfolio in one week. First we simulate the term structure starting in one week using (33). For each simulation, we find the market value of all instruments in the portfolio. By assigning equal probability to each simulation, this gives a distribution of future market values.

We have chosen a very simple example portfolio. It consists of a forward contract for the first 6 months in year 2000 and a put option with exercise date at the last day of 1999, written on the same forward. The strike on the option is 120. Figure 7 gives the distribution in one week for the forward contract. Figure 8 gives similar information for the put option. In Fig. 9, we give the statistics for the total portfolio. The example illustrates the risk reduction effect from the option on the total portfolio.

8 Conclusions

The purpose of this paper is to derive a decision support model for professionals in the electricity market for valuation and risk management. The paper applies results and methods from finance, and incorporates the fact that electricity derivatives are written on a commodity flow rather than a bulk delivery.

The electricity derivatives market is represented by a forward price function, following from the quoted prices on traded contracts. The market uncertainty is modelled by a volatility function being a decreasing (and convex) function of time.

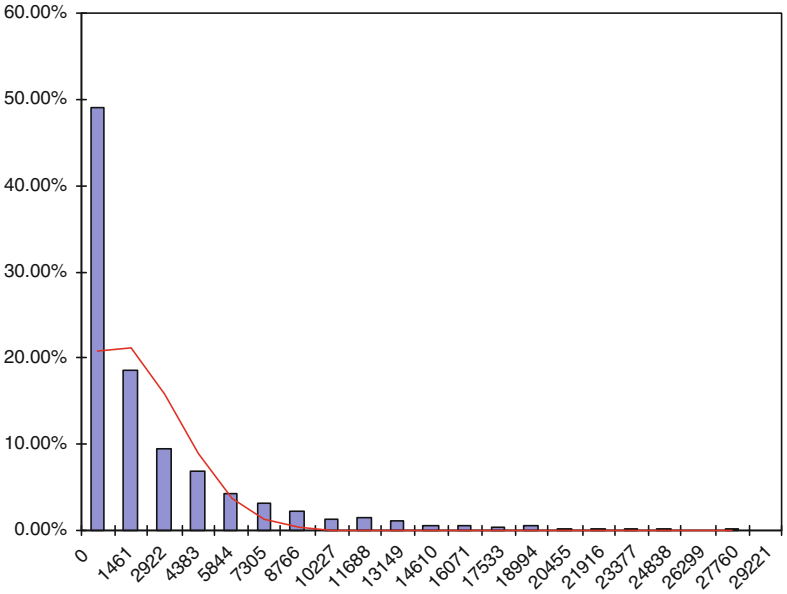


Fig. 8 Distribution for the value of a put option on the forward contract first half of 2000 in one week, strike equal 120, NOK

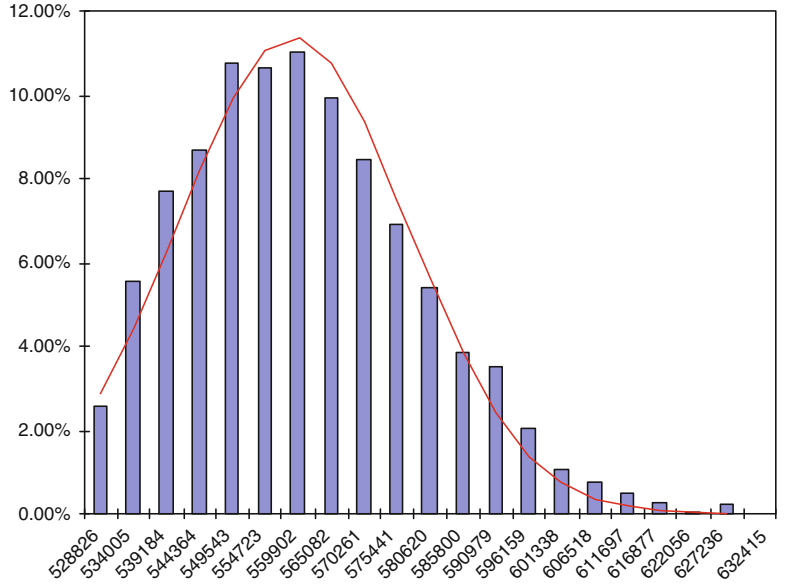


Fig. 9 Distribution for the value of a portfolio consisting of one forward and one put option, NOK

The paper presents value approximation results for "European" as well as Asian call options. The 3-factor market risk management model presented in the paper is compatible with these results and can be used for quantifying the future market risk of given portfolios (including VaR).

Acknowledgements We thank the editors, anonymous referees, and P.E. Manne for useful comments.

Appendix

This appendix evaluates (12) above. Define the new integration variable $x = b + \frac{1}{2}(T_2 + T_1) - s$ and the constant $\alpha = \frac{1}{2}(T_2 - T_1)$, and write the integral as

$$\mathrm{Var}_t^*\left[\ln\left(\frac{F(\tau,T_1,T_2)}{F(t,T_1,T_2}\right)\right] = \left(\frac{a}{T_2-T_1}\right)^2 \int_{X(\tau)}^{X(t)} \left(\ln\left(\frac{x+\alpha}{x-\alpha}\right)\right)^2 dx + \frac{2ac}{T_2-T_1}\int_{X(\tau)}^{X(t)} \ln\left(\frac{x+\alpha}{x-\alpha}\right) dx + c^2(\tau - t),$$

where $X(t) = b + \frac{1}{2}(T_2 + T_1) - t$ and $X(\tau) \equiv b + \frac{1}{2}(T_2 + T_1) - \tau$.

Observe that with $b > 0$ and $t < \tau \leq T_1 < T_2$, we have $x + \alpha > 0$ and $x - \alpha > 0$ for $x \in [X(\tau), X(t)]$. Now, use the following two results:[7]

$$\begin{aligned}\int \left(\ln\left(\frac{x+\alpha}{x-\alpha}\right)\right)^2 dx &= (x+\alpha)\,(\ln(x+\alpha))^2 \\ &\quad -2(x+\alpha)\ln(x+\alpha)\ln(x-\alpha) \\ &\quad +4\alpha\ln(2\alpha)\ln\left(\frac{x-\alpha}{2\alpha}\right) - 4\alpha\,\mathrm{dilog}\left(\frac{x+\alpha}{2\alpha}\right) \\ &\quad +(x-\alpha)\,(\ln(x-\alpha))^2 - 4\alpha, \\ \int \ln\left(\frac{x+\alpha}{x-\alpha}\right) dx &= (x+\alpha)\ln(x+\alpha) - (x-\alpha)\ln(x-\alpha) - 2\alpha,\end{aligned}$$

where

$$\mathrm{dilog}(x) \equiv \int_1^x \frac{\ln(s)}{1-s} ds.$$

Substitute the results into the variance expression, to obtain the desired result.

[7] It is straightforward to verify these results using the fact that

$$\frac{\partial}{\partial x}\,\mathrm{dilog}\,(x) = \frac{\ln(x)}{1-x}.$$

References

Abramowitz, M., & Stegun, I. (1972). *Handbook of mathematical functions*. New York: Dover.

Bjerksund, P. (1991). *Contingent claims evaluation when the convenience yield is stochastic: analytical results*. Discussion paper no. 1/1991, Institute of Finance and Management Science, Norwegian School of Economics and Business Administration.

Bjerksund, P., Rasmussen, H., & Stensland, G. (2000). *Valuation and risk management in the Norwegian electricity market*. Discussion paper no. 20/2000, Institute of Finance and Management Science, Norwegian School of Economics and Business Administration.

Gibson, R., & Schwartz, E. S. (1990). Stochastic convenience yield and the pricing of oil contingent claims. *Journal of Finance*, *45*, 959–976.

Harrison, M. J., & Kreps, D. (1979). Martingales and arbitrage in multiperiod security markets. *Journal of Economic Theory*, *20*, 381–408.

Harrison, J. M., & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications*, *11*, 313–316.

Heath, D., Jarrow, R. A., & Morton, A. J. (1992). Bond pricing and the term structure of interest rates. *Econometrica*, *60*, 77–105.

Hilliard, J. E., & Reis, J. (1998). Valuation of commodity futures and options under stochastic convenience yields, interest rates and jump diffusions in the spot, *Journal of Financial and Quantitative Analysis*, *33*, 61–86.

Hull, J. C. (1998). *Introduction to futures and options markets* (3rd ed., pp. 338–358). New Jersey: Prentice-Hall.

Jamshidian, F., & Fein, M. (1990). *Closed-form solutions for oil futures and European options in the Gibson–Schwartz model: a note*. Working paper, Merrill Lynch Capital Markets.

Jorion, P. (1997). *Value at risk: the new benchmark for controlling market risk*. New York: McGraw-Hill.

Kemna, A. G. Z., & Vorst, A. C. F. (1990). A pricing method for options based on average asset values. *Journal of Banking and Finance*, *14*, 113–129.

Lindell, A., & Raab, M. (2008). *Replacement of payoff function and hedging with a correlated asset in the case of a real power option*. Research report 2008:4, Mathematical Statistics, Stockholm University.

Miltersen, K. R., & Schwartz, E. S. (1998). Pricing of options on commodity futures with stochastic term structures of convenience yields and interest rates. *Journal of Financial and Quantitative Analysis*, *33*, 33–59.

Schwartz, E. S. (1997). The stochastic behavior of commodity prices: implications for valuation and hedging. *Journal of Finance*, *52*, 923–973.

Stochastic Programming Models for Short-Term Power Generation Scheduling and Bidding

Trine Krogh Kristoffersen and Stein-Erik Fleten

Abstract We provide an overview of stochastic programming models in short-term power generation scheduling and bidding. Special emphasis is placed on the development prompted by the restructuring of the electricity sector.

1 Introduction

Traditional applications of stochastic programming to power operation planning represent a well-developed research area. The setting is based on centralized decision-making and regulated markets in which many local producers enjoy monopoly. However, the area of applications within the new environment of decentralized decision-making, deregulated markets and competition continues to develop.

This chapter aims to present the development in stochastic programming models for short-term power generation scheduling and bidding prompted by the restructuring of the electricity sector. Although many models apply in general, the starting point is Nordic electricity producers participating in the Nordic electricity market. We seek to explain how the models of the traditional setting can be adapted to the new environment and how some newer models become highly relevant. To discuss such models from a practical viewpoint, we include a number of applications, consider computational aspects such as decomposition potential and introduce the most common solution approaches.

The outline of the chapter is as follows. A short introduction to the stochastic programming framework is given in Sect. 2. Sections 3 and 4 are confined to power

T.K. Kristoffersen (✉)
Risø National Laboratory for Sustainable Energy, Technical University of Denmark, 4000 Roskilde, Denmark
e-mail: trkr@risoe.dk

S.-E. Fleten
Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway
e-mail: stein-erik.fleten@iot.ntnu.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_12,
© Springer-Verlag Berlin Heidelberg 2010

generation scheduling in regulated and deregulated markets, respectively, whereas Sect. 5 presents a selection of solution approaches to such problems. Section 6 discusses the bidding problems that have arisen with the restructured markets along with approaches for solving them.

For a more general survey on stochastic programming problems in energy, see Wallace and Fleten (2003) who consider both short-term, medium-term and long-term problems.

2 The Stochastic Programming Framework

The models of the following sections are multi-stage stochastic programming models. Due to only limited information on some data, decisions are made under uncertainty. When decisions are made without anticipating future realizations of uncertain data, decisions are partitioned into stages according to the information flow. Hence, when the realization of uncertain data is only gradually revealed, decisions are made dynamically. Uncertainty is often described by a finite set of scenarios and corresponding scenario probabilities, a scenario being a path of realizations over time. Scenarios can be generated, for example, from historical data, by the matching of statistical properties, Høyland and Wallace (2001); Høyland et al. (2003), or by sampling from statistical models, Shapiro (2003); Pennanen (2005). To ensure that the same information always induces the same decisions, uncertainty can be represented by a so-called scenario tree in which scenarios are clustered so that branching occurs with the arrival of new information and decisions are taken at the nodes of the tree. Hence, the scenario tree is based on a set of nodes $\mathcal{N}$. Apart from the root node, all nodes have an ascendant node and a set of descendant nodes. For node n, the immediate ascending node is termed n_{-1} with the transition probability $\pi^{n/n_{-1}}$, that is, the probability that n is the descendant of n_{-1}. The probabilities of the nodes are then given recursively by $\pi^1 = 1$ and $\pi^n = \pi^{n/n_{-1}}\pi^{n_{-1}}, n > 1$. A scenario tree is illustrated in Fig. 1. For more on the notation of scenario trees in stochastic programming, see Römisch and Schultz (2001).

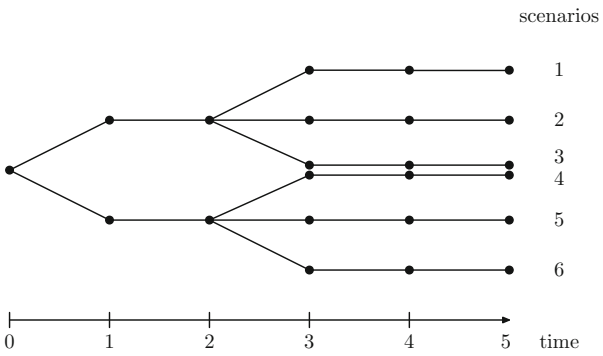


Fig. 1 Multi-stage scenario tree

3 Power Generation Scheduling in Regulated Markets

Due to centralized decision-making, power generation scheduling within the setting of regulated markets may concern several producers that are voluntarily or legally coordinated. By social welfare arguments, the models rest upon cost minimization subject to demand satisfaction.

3.1 Thermal Unit Commitment and Hydro-Thermal Scheduling

As a starting point, we address the thermal unit commitment problem from the perspective of a regulated utility. The stochastic programming problem consists in the dynamic scheduling of start-ups, shut-downs and operation levels of the thermal units such as to minimize expected costs of meeting an uncertain demand that is only revealed gradually.

To formalize this, let $\mathcal{I}$ index the thermal units. Denote the on/off-decisions of the units $u_i^n, i \in \mathcal{I} \in \{0,1\}, n \in \mathcal{N}$ and let the $p_i^n \in \mathbb{R}_+, i \in \mathcal{I}, n \in \mathcal{N}$ represent the corresponding operation levels. Due to the node dependency decisions can adapt dynamically to realized demand. Expected costs account for future operational fuel costs and start-up costs. These costs are modeled by the functions $FC(\cdot)$ and $SC(\cdot)$ in (1). The functions are typically approximated by piecewise linear functions in order to obtain a mixed-integer programming formulation. Generation is subject to upper and lower bounds, $p_i^{min}, p_i^{max}, i \in \mathcal{I}$, see (2). In order to prevent thermal stress and high maintenance costs, minimum up- and down-time constraints apply. These are given by (3), where $\tau_i^{min}, \tau_i^{max}, i \in \mathcal{I}$ denote the minimum up- and down-times. In addition, must-on/-off constraints can be included. A regulated utility is forced to satisfy demand by means of generation only which leads to the constraints (4) with demand denoted by $d^n, n \in \mathcal{N}$. The utility may have similar reserve responsibilities, which induces the constraints (5) with reserve demand denoted by $r^n, n \in \mathcal{N}$. Reserves ensure excess system capacity in the case of failure. The necessity of including reserve constraints in a stochastic model depends on whether failures are explicitly incorporated in the scenarios, for example, as demand peaks corresponding to the capacities of disrupted units. Thus, the thermal unit commitment problem is

$$\min \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} \pi^n \left(FC_i(p_i^n, u_i^n) + SC_i(u_i^n, u_i^{n-1}) \right) \tag{1}$$

$$\text{s.t. } u_i^n p_i^{min} \leq p_i^n \leq u_i^n p_i^{max}, \quad i \in \mathcal{I}, n \in \mathcal{N} \tag{2}$$

$$u_i^{n-\tau} - u_i^{n-(\tau+1)} \leq u_i^n, \quad \tau = 1, \ldots, \tau_i^{max} - 1, i \in \mathcal{I}, n \in \mathcal{N} \tag{3}$$

$$u_i^{n-(\tau+1)} - u_i^{n-\tau} \leq 1 - u_i^n, \quad \tau = 1, \ldots, \tau_i^{min} - 1, i \in \mathcal{I}, n \in \mathcal{N}$$

$$\sum_{i \in \mathcal{I}} p_i^n \geq d^n, \quad n \in \mathcal{N} \tag{4}$$

$$\sum_{i \in \mathcal{I}} (u_i^n \, p_i^{max} - p_i^n) \geq r^n, \quad n \in \mathcal{N} \tag{5}$$

$$p_i^n \geq 0, u_i^n \in \{0, 1\}, \quad i \in \mathcal{I}, n \in \mathcal{N}.$$

Early attempts to formulate mathematical programming models in power operation planning date back to Kirchmayer (1958) who considered the operation of power systems in general, and Baldwin et al. (1959) who studied the shut-down of generating units. A deterministic unit commitment model was found in, for example, Muckstadt and Koenig (1977), before finally Fahd and Sheble (1994) provided an overview of the early unit commitment formulations. With the progress in stochastic programming, the stochastic extensions of the deterministic operation planning problems began to take form with, for example, Bunn and Paschentis (1986), who investigated economic dispatch of electricity under demand uncertainty, as well as Takriti et al. (1996); Carpentier et al. (1996), who considered the unit commitment problem when subjected to uncertain demand, unit failures, etc. The model above is similar in spirit.

Thermal power production can be combined with hydropower generation using pumped storage plants, the result being referred to hydro-thermal generation. Hydro-thermal scheduling consists in the planning of thermal power production with the possibility of using electricity to pump up water and store it in the reservoirs for future hydropower generation. The planning may be complicated by uncertainty in inflows to the reservoirs. The objective is therefore to minimize expected thermal operational costs as well as opportunity costs of hydropower generation that arise since the water could be saved for future use. The opportunity costs are measured as the future value of stored water. Various heuristics can be used to estimate this value, including the use of information in long-term forward contracts. As an alternative, Fosso et al. (1999) suggest to calculate marginal water values for a hydro system as derivatives of stochastic dynamic programming value functions with respect to reservoir storage levels and use these to obtain water values for each reservoir in the system. Including water values is an attempt to avoid end effects due to a finite time horizon, such as the tendency of the multi-stage stochastic programming problem to empty the reservoirs in the final stage. Short-term hydro-thermal scheduling models in multi-stage stochastic programming are found several places in the literature, see Dentcheva and Römisch (1998); Gröwe-Kuska et al. (2000, 2002); Nowak and Römisch (2000). However, as an alternative to including water values, the authors constrain the final storage levels to a certain size.

Similar in spirit to the hydro-thermal scheduling models are the two-stage unit commitment models of Carøe and Schultz (1998) and Gollmer et al. (2000). The models also seek to find a unit commitment schedule for thermal units in a hydro-thermal utility. Since, however, coal-fired units have longer start-up times than gas-burning units, the on-/off-decisions of the coal-fired units are first-stage,

whereas corresponding decisions of the gas-burning units are second-stage. For other two-stage hydro-thermal planning problems, see also Dentcheva and Römisch (1998) and Nowak et al. (2000), who assign a full schedule in the first stage and a compensation schedule in the second stage.

3.2 Hydro Scheduling

When hydro scheduling is addressed from the point of view of a regulated hydropower utility, the problem consists in the spatial distribution of water releases between different hydro reservoirs such as to satisfy a possibly uncertain electricity demand. Since direct operating costs of hydropower generation are negligible, the water releases from the reservoirs are determined by a balance between the immediate and future value of their contents. Challenges are posed by natural inflows being uncertain and reservoirs being connected by a network in which releases upstream contribute to inflows downstream, possibly with a time delay. Although formulated as medium-term hydro scheduling, Jacobs et al. (1995) present such a model. The model includes an entire network of lakes, reservoirs, water courses, tunnels, junctions and power houses. Since the hydropower system is connected to a thermal system, the major costs are those of avoiding thermal generation. A complete review on deterministic and stochastic reservoir management models is given by Yeh (1985).

The hydropower unit commitment problem introduces the scheduling of start-ups and shut-downs into the hydro scheduling problem. It follows that the overall problem consists in determining on/off-schedules and corresponding generation levels of the turbines so as to balance current costs and future water values. Most direct operating costs can be ignored so that current costs account for start-up costs only, although even hydro start-up costs are much lower than thermal start-up costs. Hydropower unit commitment has been addressed only few times in the literature, one of the few examples being Philpott et al. (2000).

4 Restructured Markets

Deregulated markets allow for decentralized decision making and the main focus of this section is therefore a single producer. For the modelling of an entire power system, guidelines are provided by, for example, Bjørndal and Jørnsten (2005). In deregulated markets, previous obligations to serve demand are replaced by the opportunity of power producers to buy and sell production of any volume. This section assumes that the producer is a price-taker, which can be justified by the size of the producer and by the number of other market participants. Nevertheless, in practice, market price manipulations may sometimes occur, for example, due to the isolation of producers and consumers caused by grid congestion. For a discussion of congestion management, see Bjørndal and Jørnsten (2001).

4.1 Thermal Unit Commitment

We will illustrate how the traditional models can be adapted to the deregulated environment using the thermal unit commitment model of Sect. 3.

With the possibility of trading in the day-ahead market and the financial markets, electricity production can be disposed of both through traditional bilateral contracts and through newer physical and financial market contracts. Moreover, power producers can purchase electricity from the markets. We now follow the lines of Wallace and Fleten (2003). Denote by $d^n, n \in \mathcal{N}$ the demand of bilateral contracts and let the variables $y^{+,n}, y^{-,n} \in \mathbb{R}_+, n \in \mathcal{N}$ represent market contracts for selling and buying, respectively. With no constraints on the market contracts, these add the flexibility necessary for production, demand, purchases and disposals to match. This leads to the equality constraint

$$\sum_{i \in \mathcal{I}} p_i^n = d^n + y^{+,n} - y^{-,n}, \quad n \in \mathcal{N}. \tag{6}$$

The expected revenues of future market disposals and purchases amount to

$$\sum_{n \in \mathcal{N}} \pi^n \rho^n (y^{+,n} - y^{-,n}), \tag{7}$$

where $\rho^n, n \in \mathcal{N}$ denote the market prices. By substitution in (7), the demand constraints (6) can be relaxed. Demand constraints may however still be present in the very short term due to, for example, day-ahead commitments, see Fleten and Kristoffersen (2008). The reserve constraints can likewise be relaxed as reserves are often the responsibility of the power system operator in a deregulated market. Relaxing these constraints, the model decouples with respect to thermal units and, thus, decision-making can be conducted on a unit basis. With the introduction of revenues from market disposals and purchases, the objective has shifted from cost minimization to profit maximization. Furthermore, a new significant source of uncertainty has come into play since market price information may be limited. The modified thermal unit commitment problem is therefore

$$\begin{aligned} \max \ & \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \pi^n \left(\rho^n p_i^n - FC_i(p_i^n, u_i^n) - SC_i(u_i^n, u_i^{n-1}) \right) \\ \text{s.t. } & (2) - -(3) \\ & p_i^n \geq 0, u_i^n \in \{0,1\}, \quad i \in \mathcal{I}, n \in \mathcal{N}. \end{aligned}$$

In line with the above, the authors Takriti et al. (2000) address the thermal unit commitment as a stochastic programming problem and include both buying and selling of electricity to a spot market in which prices are not known in advance. The buying and selling are modelled as two additional units. Bounds on these units are automatically imposed by maximum demand and supply. For an overview

of the deregulated electricity system in Norway and the corresponding long-term, medium-term and short-term models used for hydro scheduling, see Fosso et al. (1999).

5 Solution Approaches

5.1 Thermal Unit Commitment and Hydro-Thermal Scheduling

The solution approaches to stochastic power operation problems are often extensions from the deterministic case. Primal approaches rely on linear programming (LP) and LP-based branch and bound, highly supported by the advances in hardware and the development of software implementations. General purpose codes can combine the LP methodology with a variety of options for arranging the branch and bound such as branching rules, fast heuristics, etc. Most importantly, the LP-based branch and bound works with ample enrichment as long as the model is expressed in mixed-integer linear terms. The drawback of the approach is that it handles a full model which is not feasible for high-dimensional integer problems. By branching in one dimension at a time, the size of the branch and bound tree increases exponentially with the dimension. This paves the way for the following decomposition methods.

Among the dual approaches, Lagrangian relaxation has proved to be a strong tool with the algorithmic progress for solving the Lagrangian dual, the usually small duality gap and the advance of fast Lagrangian heuristics. For the application of Lagrangian relaxation in the deterministic case, see Feltenmark et al. (1997); Gollmer et al. (1999); Jörnsten et al. (1985). For a justification of the application to stochastic programming problems, we refer the reader to Rockafellar and Wets (1978); Bjørnestad et al. (1988).

To illustrate, we develop the Lagrangian dual of the multi-stage stochastic programming version of the thermal unit commitment problem of Sect. 3. The problem is nearly separable with respect to thermal units as only the constraints (4) and (5) couple different units. This property may be utilized by stochastic Lagrangian relaxation of the unit-coupling constraints. Assigning non-negative stochastic Lagrange multipliers that have the same tree structure as $d^n, r^n, n \in \mathcal{N}$, the Lagrangian is

$$\begin{aligned} L(u, p; \lambda_1, \lambda_2) := & \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} \pi^n \left(FC_i(p_i^n, u_i^n) + SC_i(u_i^n, u_i^{n-1}) \right) \\ & + \sum_{n \in \mathcal{N}} \pi^n \lambda_1^n \left(d^n - \sum_{i \in \mathcal{I}} p_i^n \right) \\ & + \sum_{n \in \mathcal{N}} \pi^n \lambda_2^n \left(r^n - \sum_{i \in \mathcal{I}} (u_i^n p_i^{max} - p_i^n) \right) \end{aligned}$$

and the corresponding dual function is given by

$$D(\lambda_1, \lambda_2) := \min_{\mathbf{u},\mathbf{p}} \{L(u, p; \lambda_1, \lambda_2) : (2) - (3)\}. \tag{8}$$

The Lagrangian dual now reads

$$\max \{D(\lambda_1, \lambda_2) : (\lambda_1, \lambda_2) \in \mathbb{R}_+^{2|\mathcal{N}|}\}. \tag{9}$$

Due to integrality restrictions, the primal problem is nonconvex. The dual problem therefore only provides a lower bound to the primal problem, although the Lagrangian relaxation provides a tighter bound than linear relaxation.

The problem (8) decomposes into single-unit subproblems. In this fashion, the dual function

$$D(\lambda_1, \lambda_2) = \sum_{i \in \mathcal{I}} D_i(\lambda_1, \lambda_2) + \sum_{n \in \mathcal{N}} \pi^n \left(\lambda_1^n d^n + \lambda_2^n r^n\right)$$

is evaluated by solving subproblems of the form

$$D_i(\lambda_1, \lambda_2) = \min_{u_i} \Bigg\{ \sum_{n \in \mathcal{N}} \pi^n \big(\min_{p_i^n} \{FC_i(p_i^n, u_i^n) - (\lambda_1^n - \lambda_2^n) p_i^n : (2)\} + SC_i(u_i^n, u_i^{n-1}) - \lambda_2^n u_i^n p_i^{max}\big) : (3) \Bigg\}.$$

Hence, the dimensionality problem of the unit commitment problem is avoided by separately solving smaller single-unit subproblems, potentially utilizing the similarities between the subproblems. The subproblems are multi-stage stochastic programming problems. For an outline of how to solve the subproblems by dynamic programming, including how to cope with minimum up- and down times, see for example, Nowak and Römisch (2000).

Since the Lagrangian dual (9) is a concave and non-differentiable problem, it was originally solved with subgradient procedures. Currently, more refined methods such as cutting plane or bundle methods have been successfully applied. An example is the proximal bundle method used in Dentcheva and Römisch (1998); Gröwe-Kuska et al. (2002); Nowak and Römisch (2000). On the basis of function evaluations and subgradient information, the method constructs a bundle of linearizations of the dual function to be maximized.

Mostly, the dual solution provided by a cutting plane or bundle method violates the demand and reserve constraints and thereby produces a duality gap. A Lagrangian heuristic is therefore used to determine a feasible and almost optimal solution of the primal problem. In most cases, the heuristic seeks to find a unit commitment solution that induces economic dispatch of the plant. The heuristic suggested in Gröwe-Kuska et al. (2002) perturbs the Lagrange multipliers such as to obtain primal feasible unit commitment schedules. By fixing binary variables

that do not change with the perturbation the size of the problem is drastically decreased. The remaining variables are switched off one at a time as long as feasibility persists.

The authors Dentcheva and Römisch (1998); Gröwe-Kuska et al. (2002); Nowak and Römisch (2000) employ the stochastic Lagrangian relaxation to the stochastic hydro-thermal scheduling problem. This prompts a decomposition into both single-unit thermal and hydro subproblems and opens the possibility of heuristics that exploit the additional flexibility of hydropower-pumped storage plants. The two-stage problems by Dentcheva and Römisch (1998); Nowak et al. (2000) are solved in the same fashion as the multi-stage problems.

Among other dual solution approaches to stochastic power operation problems is Lagrangian relaxation of the non-anticipativity constraints. An early reference on Lagrangian relaxation of the non-anticipativity constraints is Jörnsten et al. (1985). To state a few examples from power operation planning, Takriti et al. (1996, 2000) solve the multi-stage stochastic unit commitment problem by progressive hedging, whereas Carøe and Schultz (1998); Gollmer et al. (2000) solve the two-stage version of the problem by dual decomposition.

5.2 Hydro Scheduling

The application of dynamic programming to hydro scheduling is supported by the sequential structure of the decision-making process and justified by the fact that the problems have relatively few stage-coupling constraints. As an example, an application of dynamic programming to hydropower unit commitment problems is found in Philpott et al. (2000). To facilitate computations, the continuous reservoir storage and discharge levels are often discretized. However, with the discretization of the state space, the full dynamic programming approach is known to suffer from the curse of dimensionality and is able to handle only a few reservoirs. In order to restrict the state space, it has been proposed to aggregate reservoirs and power stations or to decompose the dynamic programming problem according the reservoirs.

Another way of avoiding the curse of dimensionality in multi-stage stochastic linear programs is by the application of nested Benders' decomposition. As an example Benders' decomposition has been applied by Jacobs et al. (1995) to a hydro-scheduling module of a larger stochastic hydro-thermal scheduling problem. The authors suggest a number of algorithmic enhancements such as the use of warm start bases, initial cut generation, disaggregated cuts and tree traversing strategies, all further explored in Morton (1996). The performance of the enhanced algorithm is tested on a collection of multi-stage stochastic hydro-scheduling problems. The algorithm is outperformed by general LP optimizers in the deterministic case but is preferable as the number of scenarios increase.

For further comparisons of multi-stage stochastic hydro-scheduling algorithms, the authors Archibald et al. (1996) investigate the revised simplex method, full dynamic programming with a uniform discretization of the state space, dynamic

programming decomposition according to reservoirs and nested Benders' decomposition. The authors find the nested Benders' decomposition approach to be the fastest, followed by dynamic programming decomposition, the revised simplex method and finally full dynamic programming. It should be remarked that although most results favor the nested Benders' decomposition, the approach fails to solve stochastic programs with several stages due to the explosion in the number of sub-problems. Moreover, Benders' decomposition applies only to stochastic linear programming problems and, hence, is not feasible for unit commitment problems.

The major drawbacks of both dynamic programming and nested Benders' decompositions can be avoided by stochastic dual dynamic programming proposed by Pereira and Pinto (1991). To avoid the curse of dimensionality, the dynamic programming value function is described by supporting hyperplanes as is the case in nested Benders' decomposition. To prevent the number of Benders' subproblems from exploding, smaller sets of subproblems are chosen by sampling from the set of scenarios at every stage. As a result, an upper bound can be estimated and its confidence limits can be used as a guideline for stopping the algorithm. It should be remarked, however, that stochastic dual dynamic programming is not capable of incorporating market price uncertainty.

6 Physical Market Exchange and Bidding

With the liberalization of the power markets, new problems have arisen and the need for newer models has been obvious. Of great importance are especially the problems of physical exchange and bidding in power markets.

Most bidding problems consider a number of one-period sealed auctions in an electricity market organized as a pool in which a uniform clearing price rule applies. As concerns stochastic programming models, Fleten and Pettersen (2005) propose a mixed-integer linear program for constructing bidding curves to be submitted to the Nordic day-ahead market. The model applies to a price-taking retailer who supplies to end users under both price and demand uncertainty. As the problem consists in demand-side bidding, decision-making can be made on an hourly basis. In contrast, approaching the bidding from the point of view of a hydropower producer day-ahead market exchange and hydropower generation has to be coordinated. As a result, time coupling due to, for example, reservoir balancing forces decision-making to be effected on a daily basis, cf. Fleten and Kristoffersen (2007). Similar in some respects, Nowak et al. (2005) model simultaneous power production and physical market exchange of a hydro-thermal producer who is, however, able to influence market prices. The producer is subjected to uncertainty in foreign bids.

For the modelling of bidding curves, define a bid as a price-volume pair (x, p). The problem of selecting both a price p and a volume x is nonlinear. To circumvent this problem, the continuous price range can be discretized into a finite number of fixed price points $p_1 \leq \cdots \leq p_{\mathcal{H}}$ such that possible bids are $(x_1, p_1), \ldots, (x_{\mathcal{H}}, p_{\mathcal{H}})$

where only the volumes $x_1, \ldots, x_{\mathcal{H}} \in \mathbb{R}_+$ have to be selected. Let the bidding curve be defined by the relation between volume and price, denoted by y and ρ, respectively. Nowak et al. (2005) suggest the use of hourly block bids such that the resulting bidding curve is a nondecreasing step function. In contrast, Fleten and Pettersen (2005) perform a linear interpolation between the price-volume points and construct a nondecreasing piece-wise linear bidding curve that is consistent with the rules of the Nordic day-ahead market

$$y = \begin{cases} \dfrac{\rho - p_1}{p_2 - p_1} x_2 + \dfrac{p_2 - \rho}{p_2 - p_1} x_1 & , \text{if } \; p_1 \le \rho < p_2 \\ \vdots & \\ \dfrac{\rho - p_{H-1}}{p_H - p_{H-1}} x_{H-1} + \dfrac{p_H - \rho}{p_H - p_{H-1}} x_H & , \text{if } \; p_{H-1} \le \rho \le p_H . \end{cases}$$

In Fleten and Kristoffersen (2007), the modelling of the Nordic day-ahead market includes both a piece-wise linear bidding curve and block bids of longer durations.

For further studies based on the price-taker assumption, Neame et al. (2003) consider a generator making offers into an electricity spot market under price uncertainty. Since the generator does not affect market prices, the optimal offers reflect the marginal costs of generation which is illustrated on, for example, hydro scheduling. However, Prichard and Zakeri (2003) argue that since the costs of hydropower generation include only opportunity costs of released water, there is no simple way to determine marginal costs and bid accordingly. Instead the authors suggest the use stochastic dynamic programming for deriving bidding curves. The dynamic programming approach is also used by Pritchard et al. (2005) for deriving piecewise constant bids, taking into account both both short-term and long-term effects of hydropower production. Similarly, Fleten and Steinsbø (2008) seek to obtain a balance between short-term and long-term effects by means of stochastic mixed-integer programming. Finally, in contrast to day-ahead bidding, De Ladurantaye et al. (2007) introduce a stochastic programming model for bidding into a two-hour-ahead market. Other contributions to optimizing bidding strategies for price-takers include Contreras et al. (2002); Conejo et al. (2002); Lu et al. (2004); Ni et al. (2004).

Bidding strategies for generators with market power are investigated in, for example, Anderson and Philpott (2002). Their bidding curve is modelled as a continuous parametrized curve, considering both smooth curves and the extension to step functions in the case of a finite number of bids. The clearing of the market is established in a separate network flow model through which the generator is able to influence market clearing prices and prices are further affected by random demand and random supply of competitors. To encapsulate the effect of uncertainty on the dispatch of the generator, the so-called market distribution function describes the probability of not being fully dispatched in the market and the optimal bidding curve is selected such as to maximize expected profit in terms of this market distribution function. The resulting optimization problem is a nonlinear control problem. Philpott and Schultz (2006) integrate the framework with thermal scheduling and unit commitment. The authors propose two-stage decision problems for optimizing

offers of either a single or several thermal units. In both cases, one stage consists of a unit commitment part handled by dynamic programming whereas the other stage involves a bidding part for which optimality conditions are derived.

Acknowledgments Kristoffersen acknowlegdes support from the Carlsberg Foundation in Denmark. Fleten acknowledges support from the Research Council of Norway through project 178373/S39, and recognizes the Norwegian research centre CenSES, Centre for Sustainable Energy Studies.

References

Anderson, E. J., & Philpott, A. B. (2002). Optimal offer construction in electricity markets. *Mathematics of Operations Research*, *27*(1), 82–100.

Archibald, T. W., Buchanan, C. S., McKinnon, K. I. M., & Thomas, I. C. (1996). Nested Benders decomposition and dynamic programming for reservoir optimization. *Journal of Operational Research Society*, *50*, 468–479.

Baldwin, C., Dale, K., & Dittrich, R. (1959). A study of the economic shutdown of generating units in daily dispatch. *AIEE Transactions on Power Apparatus and Systems*, *78*, 1272–1284.

Bjørndal, M., & Jørnsten, K. (2001). Zonal pricing in a deregulated energy market. *The Energy Journal, 22*(1), 51–73.

Bjørndal, M., & Jørnsten, K. (2005). The deregulated electricity market viewed as a bilevel programming problem. *Journal of Global Optimization, 33*(3), 465–475.

Bjørnestad, S., Hallefjord, Å., & Jörnsten, K. (1988). Discrete optimization under uncertainty: The scenario and policy aggregation technique. Technical report, Chr. Michelsen Institute, Bergen, Norway.

Bunn, D. W., & Paschentis, S. N. (1986). Development of a stochastic model for the economic dispatch of electric power. *European Journal of Operational Research, 27*, 179–191.

Carøe, C. C., & Schultz, R. (1998). A two-stage stochastic program for unit commitment under uncertainty in a hydro-thermal power system. Konrad-Zuse-Zentrum für Informationstechnik Berlin, Preprint SC 98-11.

Carpentier, P., Cohen, G., Culioli, J. C., & Renaud, A. (1996). Stochastic optimization of unit commitment: A new decomposition framework. *IEEE Transactions on Power Systems, 11*(2), 1067–1073.

Conejo, A. J., Nogales, F. J., & Arroyo, J. M. (2002). Price-taker bidding strategy under price uncertainty. *IEEE Transactions on Power Systems, 17*(4), 1081–1088.

Contreras, J., Candiles, O., Ignacio de la Fuente, J., & Gomez, T. (2002). A cobweb bidding model for competitive electricity markets. *IEEE Transactions on Power Systems, 17*(1), 148–153.

De Ladurantaye, D., Gendreau, M., & Potvin, J. Y. (2007). Strategic bidding for price-taker hydroelectricity producers. *IEEE Transactions on Power Systems, 22*(4), 2187–2203.

Dentcheva, D., & Römisch, W. (1998). Optimal power generation under uncertainty via stochastic programming. In *Stochastic programming methods and technical applications* (Vol. 458, pp. 22–56), *Lecture Notes in Economics and Mathematical Systems*. Springer.

Fahd, G. N., & Sheble, G. B. (1994). Unit commitment literature synopsis. *IEEE Transactions on Power Systems, 9*(1), 128–135.

Feltenmark, S., Kiwiel, K. C., & Lindberg, P. -O. (1997). Solving unit commitment problems in power production planning. In U. Zimmermann (Ed.), *Operations research proceedings* (pp. 236–241). Berlin: Springer.

Fleten, S. -E., & Kristoffersen, T. K. (2007). Stochastic programming for optimizing bidding strategies of a nordic hydropower producer. *European Journal of Operational Research, 181*, 916–928.

Fleten, S. -E., & Kristoffersen, T. K. (2008). Short-term hydropower production planning by stochastic programming. *Computers and Operations Research, 35*, 2656–2671.

Fleten, S. -E., & Pettersen, E. (2005). Constructing bidding curves for a price-taking retailer in the Norwegian electricity market. *IEEE Transactions on Power Systems, 20*(2), 701–708.

Fleten, S. -E., & Steinsbø, J. A. (2008). Spot market bidding for a Nordic hydropower producer - a stochastic approach to water values. Working paper, Norwegian University of Science and Technology.

Fosso, O. B., Gjelsvik, A., Haugstad, A., Mo, B., & Wagensteen, I. (1999). Generation scheduling in a deregulated system. The Norwegian case. *IEEE Transactions on Power Systems, 14*(1), 75–81.

Gollmer, R., Möller, A., Nowak, M. P., Römisch, W., & Schultz, R. (1999). Primal and dual methods for unit commitment in a hydro-thermal power system. In *Proceedings of the 13th power systems computation conference* (pp. 724–730). Trondheim.

Gollmer, R., Nowak, M. P., Römisch, W., & Schultz, R. (2000). Unit commitment in power generation - a basic model and some extensions. *Annals of Operations Research, 96*(1-4), 167–189.

Gröwe-Kuska, N., Kiwiel, K. C., Nowak, M. P., Römisch, W., & Wegner, I. (2002). Power management in a hydro-thermal system under uncertainty by Lagrangian relaxation. In A. Ruszczyński & C. Greengard (Eds.), *Decision making under uncertainty: energy and power* (Vol. 128, pp. 39–70), *IMA volumes in mathematics and its applications*. New York: Springer.

Gröwe-Kuska, N., Römish, W., & Nowak, M. P. (2000). Power management under uncertainty by Lagrangian relaxation. In *Proceedings of the 6th international conference probabilistic methods applied to power systems PMAPS 2000* (Vol. 2). INESC Porto.

Høyland, K., Kaut, M., & Wallace, S. W. (2003). A heuristic for moment-matching scenario generation. *Computational Optimization and Applications, 24*(2–3), 169–185.

Høyland, K., & Wallace, S. W. (2001). Generating scenario trees for multistage decision problems. *Management Science, 47*(2), 295–307.

Jacobs, J., Freeman, G., Grygier, J., Morton, D., Schultz, G., Staschus, K., & Stedinger, J. (1995). SOCRATES: a system for scheduling hydroelectric generation under uncertainty. *Annals of Operations Research, 59*, 99–133. Models for planning under uncertainty.

Jörnsten, K. O., Näsberg, M., & Smeds, P. A. (1985). Variable splitting–a new Lagrangean relaxation approach to some math. program. models. Technical report, Linköping Institute of Technology, Department of Mathematics.

Kirchmayer, L. K. (1958). *Economic operation of power systems*. New York: Wiley.

Lu, N., Chow, J. H., & Desrochers, A. A. (2004). Pumped-storage hydro-turbine bidding strategies in a competitive electricity market. *IEEE Transactions on Power Systems, 19*(2), 834–841.

Morton, D. P. (1996). An enhanced decomposition algortihm for multistage stochastic hydroelectric scheduling. *Annals of Operations Research, 64*, 211–235.

Muckstadt, J. A., & Koenig, S. A. (1977). An application of Lagrangian relaxation to scheduling in power generation systems. *Operations Research, 25*(3), 387–403.

Neame, P. J., Philpott, A. B., & Pritchard, G. (2003). Offer stack optimization in electricity pool markets. *Operations Research, 51*(3), 397–408.

Ni, E., Luh, P. B., & Rourke, S. (2004). Optimal integrated generation bidding and scheduling with risk management under a deregulated power market. *IEEE Transactions on Power Systems, 19*(1), 600–609.

Nowak, M. P., Nürnberg, R., Römisch, W., Schultz, R., & Westphalen, M. (2000). Stochastic programming for power production and trading under uncertainty. Submitted.

Nowak, M. P., & Römisch, W. (2000). Stochastic Lagrangian relaxation applied to power scheduling in a hydro-thermal system under uncertainty. *Annals of Operations Research, 100*(1-4), 251–272.

Nowak, M. P., Schultz, R., & Westphalen, M. (2005). A stochastic integer programming model for incorporating day-ahead trading of electricity into hydro-thermal unit commitment. *Optimization and Engineering, 6*, 163–176.

Pennanen, T. (2005). Epi-convergent discretizations of multistage stochastic programs. *Mathematics of Operations Research, 30*(1), 245–256.

Pereira, M. V. F., & Pinto, L. M. V. G. (1991). Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming, 52*, 359–375.

Philpott, A., & Schultz, R. (2006) Unit commitment in electricity pool markets. *Mathematical Programming, Series B*(108), 313–337.

Philpott, A. B., Craddock, M., & Waterer, H. (2000). Hydro-electric unit commitment subject to uncertain demand. *European Journal of Operations Research, 125*, 410–424.

Prichard, G., & Zakeri, G. (2003). Market offering strategies for hydroelectric generators. *Operations Research, 51*(3), 602–612.

Pritchard, G., Philpott, A. B., & Neame, P. J. (2005). Hydroelectric reservoir optimization in a pool market. *Mathematical Programming, 103*(3), 445–461.

Rockafellar, R. T., & Wets, R. J. -B. (1978). The optimal recourse problem in discrete time: l^1-multipliers in discrete time. *SIAM Journal on Control and Optimization, 16*(1), 16–36.

Römisch, W., & Schultz, R. (2001). Multistage stochastic integer programs: An introduction. In M. Grötchel, S. O. Krumke, & J. Rambau (Eds.), *Online optimization of large scale systems* (pp. 581–600). Berlin: Springer.

Shapiro, A. (2003). Statistical inference of multistage stochastic programming problems. *Mathematical methods of Operations Research, 58*(1), 57–68.

Takriti, S., Birge, J. R., & Long, E. (1996). A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems, 11*(3), 1497–1508.

Takriti, S., Krasenbrink, B., & Wu, L. S. -Y. (2000). Incorporating fuel constraints and electricity spot prices into the stochastic unit commitment problem. *Operations Research, 48*(2), 268–280.

Wallace, S. W., & Fleten, S. -E. (2003). Stochastic programming models in energy. In A. Ruszczyński & A. Shapiro (Eds.), *Stochastic programming* (Vol. 10, pp. 637–677) *Handbooks in operations research and management science*. Elsevier, North-Holland.

Yeh, W. W. -G. (1985). Reservoir management and operation models: a state of the art review. *Water Resources Research, 21*(12), 1797–1818.

Optimization of Fuel Contract Management and Maintenance Scheduling for Thermal Plants in Hydro-based Power Systems

Raphael Martins Chabar, Sergio Granville, Mario Veiga F. Pereira, and Niko A. Iliadis

Abstract The objective of this work is to present a decision support system that determines the optimal dispatch strategy of thermal power plants while considering the particular specifications of fuel supply agreements, such as take-or-pay and make-up clauses. Opportunities for energy purchase and selling at the spot market as well as a detailed modeling of the power plant (maintenance cycles, influence of temperature, etc.) are also considered during the optimization. In an integrated approach, the model also determines the plants' optimal schedule for maintenance. Since decisions in a stage have an impact in the future stages, the problem is time-coupled with a multi-stage framework. Moreover, the main driver for the decision-making is the energy spot price, which is unknown in the future and is modeled in this tool through user-defined scenarios. Therefore, the calculation of the optimal dispatch strategy is modeled as a decision under uncertainty problem, where at each stage the objective is to determine the optimal operation strategy that maximizes the total revenues taking into account the constraints and characteristics of the fuel supply contract. The methodology applied is a hybrid Stochastic Dual Dynamic Programming (SDDP)/Stochastic Dynamic Programming (SDP). Examples and case studies will be analyzed for the Brazilian system.

1 Introduction

As widely discussed in the literature (Chabar 2005; Granville et al. 2003a), spot price volatility in thermal based systems results mostly from fluctuations in the demand, from forced outages of equipment, and from fluctuations in fuel prices.

R.M. Chabar and S. Granville
PSR, Rio de Janeiro, RJ, Brazil
e-mail: chabar@psr-inc.com; granville@psr-inc.com

M.V.F. Pereira
PSR, Rio de Janeiro, Brasil
e-mail: mario@psr-inc.com

N.A. Iliadis
EnerCoRD, Athens, Greece
e-mail: niko.iliadis@enercord.com

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_13,
© Springer-Verlag Berlin Heidelberg 2010

Hydro dominated systems, such as the Brazilian system, have relatively small short-term volatility, but high medium-long term volatility. The reason for the reduced short-term volatility is that the hydro plants can easily transfer the energy from off-peak to peak hours. In other words, these plants usually have more than enough capacity to modulate the peak supply and thus avoid the main causes of short-term volatility. This leads to an equalization of spot prices in the short run.

Hydro systems are designed to ensure the load to be supplied even under adverse hydrological conditions. However, these conditions do not frequently occur – which is the reason for medium term volatility. As a consequence, the system has excess of energy during most of the time. This surplus allows the demand to be met without using any thermal resource, resulting in very low spot prices. Nevertheless, whenever a drought period occurs, hydro plants may turn to be unable to supply alone the system demand and as a consequence spot prices increase rapidly, and can even reach the system's rationing cost. Due to the high storage capacity of water reservoirs, these low cost periods can become long, interspersed with very high cost periods resulting from droughts. For example, Fig. 1 below shows the observed spot prices in the Brazilian Wholesale Energy Market from 2000 until early 2005.

In this context, it is known that a thermal plant that generates only when spot prices are high can meet its contract obligations with a low effective operating cost. Hence, operational flexibility is a very attractive characteristic for thermal plants in hydro-based systems.

Nevertheless, this operational flexibility, along with the low diversification of fuel markets, is in opposition with the constraints that fuel producers have. The latter participants have high fixed costs due to capital expenditures in developing production and transportation infrastructure, such as investments in gas fields (in case

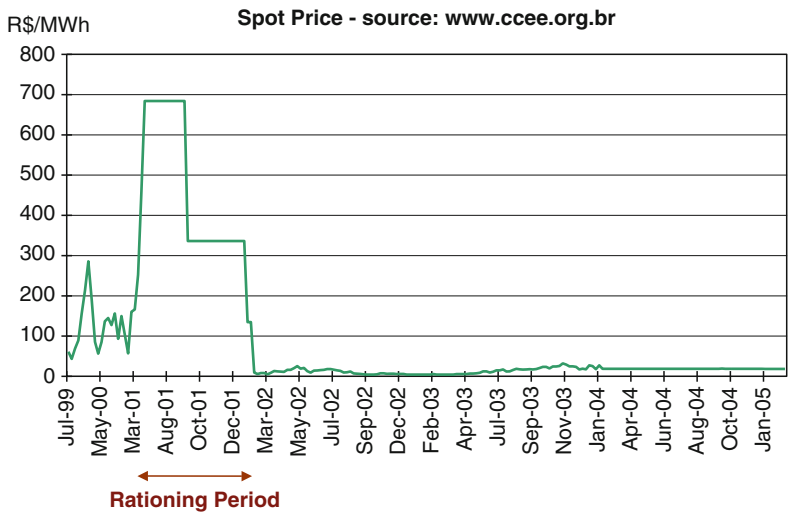


Fig. 1 Brazilian spot prices

of gas producers) and pipelines. Operational flexibility introduces high volatility to the producer's cash flow. As a consequence, fuel supply agreements are structured over take-or-pay (ToP) clauses. These are just financial agreements to reduce the volatility of the fuel producer's revenues and are not necessarily associated to consumption obligation. The ToP clauses impose an anticipated purchase of a minimum amount of fuel (on a daily, monthly, and/or yearly basis), independently of its consumption. The amount of fuel bought but not consumed is virtually "stored" for a pre-set period. During this period, the fuel can be recovered by the plant. This is known as make-up clause.

In the Brazilian power system, thermal generators declare to the Independent System Operator (ISO) their variable operating costs ($/MWh), availability (MW), and must-run generation (MW). Based on the declarations of all agents, the ISO carries out a least-cost hydrothermal scheduling of the system aiming on meeting the system demand with the lowest possible operating cost (Pereira and Pinto 1984, 1985; Pereira et al. 1998). This approach is adopted in many other countries, such as Peru and Ecuador.

Therefore, in order to profit the most from its operational flexibility considering the fuel supply agreements constraints, the generator has to develop an operation strategy that optimizes the management of the fuel contract use. Instead of firing at each stage an amount of fuel corresponding to the ToP clause, as if this clause was physical, it may be more attractive to have its dispatch reduced during low price periods and store the paid but not fired fuel for future use when spot prices are high.

Another relevant aspect consists in the definition of a maintenance schedule for the plant since all maintenance is associated to a high fixed execution cost. The less the plant is dispatched, the latter – and also the fewer – it will have to stop for maintenance, resulting in lower costs. Furthermore, the best strategy will be to schedule the maintenances for the periods with low spot prices. Due to this interdependency between dispatch strategy and maintenance scheduling, these factors should be analyzed jointly.

The objective of this work is to present a tool that determines the optimal dispatch strategy for a power plant taking into account:

- The fuel supply agreement with take-or-pay and make-up conditions,
- Upsides opportunities of trading in the spot market,
- Plant's detailed operating characteristics such as
  - Maintenance cycles
  - Power output
- Conversion rate variations due to temperature and degradation with lifetime,
- Changes in configuration over time such as
  - Changes in the number of generation units
  - Changes in the available power
- Variable and fixed costs,
- Energy sale contracts, etc.

The model also determines the optimal maintenance schedule for each generation unit individually regarding each unit's elapsed effective operating time and the maximum allowed operating time before each maintenance cycle, considering maintenance execution costs. The methodology and examples of this work will be presented using a combined cycle natural gas power plant (CCGT). The proposed approach can be easily extended for any other technology whose fuel supply agreement has take-or-pay and make-up clauses.

This work is organized as follows: Sect. 2 describes the general aspects of fuel supply agreements, maintenance scheduling and discusses decision under uncertainty aspects. Section 3 presents the general aspects of the proposed methodology and describes the tool. Examples and case studies are presented in Sect. 4. Section 5 presents the comparison between the Brazilian and European markets and the adaptations of the model for Europe. Section 6 states all of the conclusions of this work.

2 General Aspects: Fuel Supply Agreements, Maintenances, and Decision Under Uncertainty

2.1 Fuel Supply Agreements: Characteristics

A standard fuel supply agreement (FSA) establishes a volume (maximum) of fuel (in case of natural gas, it is defined in MMm^3/day) that can be withdrawn by the producer for daily consumption, besides the commodity fuel price per unit of consumption (in \$/MMBTU). If ToP clauses are applied, then, from this contracted monthly maximum amount, the generator is obliged to purchase (but not necessarily to consume) $X\%$. If, in the end of the year, the annual consumption is lower than $Y\%$ of the total annual volume of the FSA, then the plant must purchase the difference. These are the monthly and annual ToP clauses, respectively. For example, the standard FSA observed in the Brazilian power system imposes clauses of $X = 56\%$ and $Y = 70\%$. The fuel already purchased but not consumed is "virtually" stored for a period of N years (in the case of Brazil, $N = 7$). This means that the fuel can be recovered by the plant at every moment (make-up clause), respecting the maximum amount that can be daily withdrawn. These arrangements can differ from fuel to fuel.

Besides the ToP clauses on the commodity, there are (usually) ship-or-pay (SoP) clauses referring to the use of the pipeline: in the contract it is specified that the price to be paid for the transportation of each unit of fuel (in case of natural gas, it is expressed in \$/MMBTU). The thermal plant has to pay a monthly minimum of $Z\%$ of the amount contracted for the use of the transportation structure. In Brazil, $Z = 95\%$ is the typical SoP value for natural gas agreements. As opposed to the ToP on the commodity, the SoP does not have any make-up clause (the amount paid refers to the pipeline use in that specific month and cannot be deducted from future transportation if it is not used in the current month). Since there is no make-up on the transportation parcel, the SoP is seen by the generator as a fixed cost.

Finally, there is still a clause on the FSA referring to the margin of the local distribution company of natural gas, which is a fee for the use of local distribution transportation network. In case of Brazil, this clause is identical to the SoP, with a payment referring to 70% of the total volume contracted, but with a price per unit transported usually lower.

2.2 Fuel Opportunity Cost

The generator must manage the use of the contracted natural gas respecting all aforementioned clauses that are part of the FSA. As observed previously, a straightforward operation strategy of the generator is to declare a must-run generation to the ISO in order to consume all the time exactly the 70% ToP of the FSA. In this way, the generator meets, without storage, its monthly/annual requirements of ToP and local distribution margin. Under this strategy, during high spot price periods, the generator will have some additional costs of buying extra fuel (besides the amount corresponding to 70% ToP) to have its plant dispatched at maximum capacity. However, other operation policies, taking into account the characteristics and flexibilities of the FSA, may be more attractive for the generator. In those policies, the power plant would change from a fixed must-run generation full-time to a more flexible generation scheduling. This strategy will manage the use and storage of the natural gas in a more efficient way and this will reduce the purchases of extra fuel, since the fuel will be stored from low price periods to high ones.

In order to determine this strategy, the plant must determine the opportunity cost of the natural gas (whose purchase is mandatory, but not its consumption) during all stages. This opportunity cost will reveal whether it is more profitable to consume the gas today or store it for the future.

2.3 Decision Under Uncertainty

The dispatch decision and consequently the use of natural gas, depends on the spot price trajectory. During low price periods the generator prefers to meet its contracts of energy supply buying energy in the spot market and storing the gas for future use. On the other hand, during high price periods the generator will attend its demand with a lower cost than the spot market price and if the production is greater that the contracted energy, the plant can sell this excess in the spot market, earning an additional profit. As the future spot prices are unknown, the problem becomes a decision making under uncertainty. The uncertainty in prices is modeled using scenarios that represent possible trajectories of the price making the optimization problem multiperiod and stochastic (large-scale optimization problem). Considering spot prices scenarios as input data to the model, we assume that the production of the thermal plant does not affect the prices in the spot market (price taker without market

power). Such an assumption is reasonable when the installed capacity of the plant is significantly lower than the total system capacity (case for individual thermal plants in the Brazilian power system).

In hydro power systems, the spot prices in a period are highly correlated with the ones of the following period. This is due, mainly, to the time correlation of the water natural inflows of the system (inflow temporal persistence), meaning that given an observed "low" price in a period, it is likely to observe also a "low" (or "not high") price in the next period. This autocorrelation of spot prices can be modeled using a conditional probability distribution for one stage to the next one. For each possible spot price in a given period, there is associated a set of conditional probabilities that represent the evolution possibilities from this spot price state (scenario) to another in the following period. This random process can be modeled via a Markov Chain. In this way, based on the price scenarios we estimate the matrix of transition probabilities from every price state (conditioned) in stage t to every price state in period $t + 1$. Details on this Markov Chain modeling and estimation can be found in (Chabar 2005; Wallace and Fleten 2003; Flatabo et al. 1998; Gjelsvik et al. 1999).

2.4 Maintenance Scheduling

Another important aspect in the definition of a plant's operation strategy is the scheduling of maintenance. Each different type of cycle of maintenance is associated with a direct cost. As the dispatch of a flexible plant in the Brazilian system is reduced, the needs for maintenance are also lower. It is interesting to schedule maintenances to months where spot prices are low. Thus, the joint optimization of maintenance and management of the fuel contract is essential to define the optimal dispatch for the company.

2.5 Computational Model

As the modeling of hydro plants' reservoirs (Granville et al. 2003b; Wallace and Fleten 2003; Flatabo et al. 1998; Gjelsvik et al. 1999), the mechanism of virtual storage of fuel due to the monthly and annual ToP clauses of the FSA can be modeled using two fictitious reservoirs. One reservoir, called A, where all natural gas not consumed from the monthly ToP is stored. Another reservoir, called B, where the difference between the annual ToP amount and the sum of all monthly ToPs of a year is stored. The scheme of fuel supply/storage is illustrated in Fig. 2. The percentages used are 56% for monthly ToP and 70% for annual ToP.

The $0.56 \times CT_t$ Hm^3 (where CT_t is the volume of gas established in the contract for month t), corresponding to the clause of monthly ToP can be directly consumed ($CToP_t$), partially or totally, and/or be stored in the reservoir A (ARM_t).

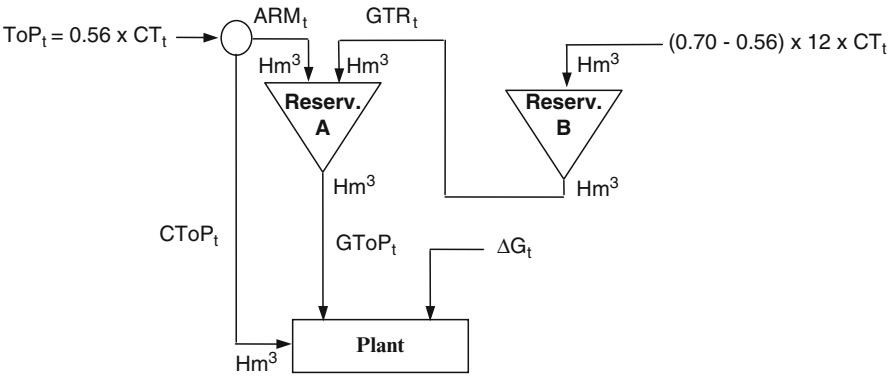


Fig. 2 Fuel supply/storage scheme

This amount is charged at the price of the commodity in period t (PGC_t), as defined in the contract. However, the recovery of the gas from reservoir A ($GToP_t$), already paid at the moment of its storage, is only allowed if the percentage of the monthly ToP has been totally consumed.

In turn, the complementary amount of the annual ToP, (difference between the 70% of the annual contract and the sum of the monthly 56%) is represented as follows: in the first month (beginning of the month) of each contractual year, there is a supply of gas to reservoir B equal to the total difference between 70% and 56% referring to the whole year. There is no payment for this initial supply. At each period t, the generator can transfer gas from reservoir B to A. This transfer (GTR_t) is charged at the PGC_t price and corresponds to "extra" monthly ToP gas purchase.

In the last month (beginning of the month) of each contractual year, all the amount of gas in B (if there is any) is transferred to reservoir A and is charged at the PGC_t price. By this operation, reservoir B is completely emptied, fulfilling the condition of the yearly ToP clause.

Moreover, in case the plant wants to produce more energy and both reservoirs A and B are empty in a given stage and all the monthly ToP of this stage has already been consumed then additional gas (ΔG_t) can be bought at an established price. The latter must occur respecting the maximum limit of monthly gas consumption established in the FSA.

The fuel stored in reservoir A has a limited period for recover (make-up clause). Due to that, an additional structure is used to "track" the storage date of all gas that enters reservoir A. For each portion stored in reservoir A, when the number of stored years reaches the limit duration, this portion is discarded ($GDesc_t$). Clearly, when an amount of gas is recovered from reservoir A, this additional structure is updated, always discarding the gas from the oldest to newest.

Similar to the ToP clauses management, the maintenance schedule can also be modeled by reservoirs. The maintenances are characterized by cycles (that correspond to a type of maintenance) and each cycle has a direct cost. A specific maintenance is necessary when a unit reaches the number of hours of operation

Table 1 Example of maintenance specifications

Cycle	Frequency (h)	Average duration (Days)	Cost (millions R$)
Combustor	8,000	7	3.5
Hot path circuit	24,000	14	10
Major maintenance	48,000	21	20

of a cycle. An example of maintenance cycles for generation units are depicted on Table 1. If an interruption for maintenance occurs on a high spot price period, the plant revenue is constrained. As a result, it is interesting to move the maintenance interruptions to low spot price periods; anticipating them even though the number of hours on a cycle has not been reached yet. The maintenance interruption can be anticipated but not postponed.

For modeling purpose, there is one reservoir of remaining hours of operation for each generation unit and each maintenance cycle. For instance, for a case with 3 cycles and 3 units, 9 reservoirs are used. These reservoirs have different capacities that correspond to the cycle frequency that they model. In the beginning of the study, for each generation unit, these reservoirs are filled with the amount of remaining hours of operation until next maintenance. As the unit operates, all reservoirs (all cycles) for that unit are reduced by the quantity of the elapsed hours. In case of a maintenance, either resulting from an anticipation or not, the unit stops during the service period, the plant "pays" the maintenance cost and the reservoir corresponding to that executed cycle is totally filled again.

The solution procedure for a stochastic multi-stage optimization problem involving reservoir modeling is well-known in literature. It has been intensively applied on hydro-thermal dispatch problems (Granville et al. 2003b; Wallace and Fleten 2003; Pereira and Pinto 1984, 1985; Pereira et al. 1998). Due to the exponential dimensionality of the problem, a hybrid Stochastic Dual Dynamic Programming (SDDP)/Stochastic Dynamic Programming (SDP) approach is appropriate to find a solution on an acceptable computational time (Pereira and Pinto 1984, 1985; Pereira et al. 1998). The details of the hybrid SDDP/SDP methodology applied are beyond the scope of this article.

In terms of dynamic programming, for a given stage and for a given price, the problem is formulated as (1):

$$\begin{aligned} FBF_t^k\left(VA_t, VB_t, \left\{VH_t^{i,j}, i = 1 \ldots N\ unit, j = 1 \ldots N\ man\right\}, \pi_t^k\right) \\ = \max RI_t + \sum_{s=1}^{s}\Big[p_{t+1}(k,s) \times FBF_{t+1}^s\left(VA_{t+1}, VB_{t+1}, \left\{VH_{t+1}^{i,j}, i\right.\right. \\ \left.\left. = 1 \ldots N\ unit, j = 1 \ldots N\ man\right\}, \pi_{t+1}^k\right)\Big] \end{aligned} \tag{1}$$

Subject to

- Fictitious reservoir balance constraints (reservoir A, B and for maintenance remaining hours), as:

$$\begin{aligned} VA_{t+1} &= VA_t + ARM_t - GToP_t + GTR_t - GDesc_t \\ VB_{t+1} &= VB_t - GTR_t \\ VH_{t+1}^{i,j} &= VH_t^{i,j} \times \left(1 - X_t^{i,j}\right) + \overline{VH}^j \times \left(X_t^{i,j}\right) - \gamma \times EG_t^j \; : \\ &\qquad \textit{for all unit i and all maintenance j} \end{aligned}$$

- Gas consumption priority: 1st $CToP_t$ is consumed, then $GToP_t$ and finally ΔG_t;
- Accounting for the immediate revenue in stage t (RI_t), which includes the financial result in the spot market, the revenues with energy supply contracts, the fuel payments (ToPs, SoP and the margin for the Distribution company (DisCo), fixed and variable operating costs and expenses with the execution of maintenances;
- Transformation of gas to energy:

$$\frac{\sum_{i=1}^{N\,unit} \varphi_t^i \times EG_t^i}{H_c} CToP_t + GToP_t + \Delta G_t$$

- Constraint of maximum gas consumption;
- Bound on the number of concurrent maintenances of the same type of cycle that may take place in a given month;
- Bound on the generation due to programmed maintenances and unexpected failure of units;
- Constraints related to the mechanism implemented for the modeling of the FSA conditions, such as the transfer of all fuel from reservoir B to A in the end of each year (clause of complement of the annual ToP) and the supply of gas in reservoir B in the beginning of each year;
- Variation of the maximum generation capacity with the environmental temperature, power degradation and efficiency (conversion of gas to energy) of the generation units with the elapsed hours of operation;
- Incidence of tariffs and taxes.

Where

- VA_t is the volume of gas in reservoir A, VB_t is the volume of gas in reservoir B and $VH_t^{i,j}$ is the "volume" of remaining hours of operation that the unit i has until the next maintenance of cycle j (Nunit is the number of generation units and Nman is the number maintenance cycles);
- π_t^s is the spot price in stage t and scenario s;
- $p_{t+1}(k, s)$ is the transition probability of spot price of scenario k in stage t (known value in stage t) to the spot price of scenario s in stage $t + 1$ (conditional transition probability of price).

- $X_t^{i,j}$ is the binary decision variable associated to the schedule of maintenance of cycle j for the unit i in stage t.
- $\overline{\mathrm{VH}}^j$ is the maximum capacity of the reservoir of remaining hours of operation until the next maintenance of cycle j (e.g., 8,000 h, 24,000 h, etc.).
- $\mathrm{EG}_\mathrm{t}^\mathrm{i}$ is the energy produced in MWh by the unit i in stage t.
- γ is the inverse of the power of each unit; φ_t^i is the conversion factor from MMBTU to MWh of unit i (efficiency); H_c is the heat rate of the gas (in $\mathrm{BTU/m^3}$).

This Mixed Integer Linear Programming (MILP) problem is solved for each time stage (month) and spot price scenario in an iterative process until the convergence of the SDDP/SDP algorithm (details of the algorithm are found in Chabar 2005). In the problem, the immediate revenue at each stage is maximized taking into consideration the probable future revenues that may occur given (conditioned to) the current spot price. The future revenues are represented by the future benefit functions (FBF) also known as cost-to-go functions, which translate the opportunity cost of storage (storage of fuel in reservoirs A and B, and "storage" of hours in the remaining hours in the maintenance reservoirs). In this way, at each stage and for each state of the spot price, the optimal solution is the one that maximizes the sum of the immediate revenue plus the expected value of the future revenues. The FBF of stage t and price level k describes the total future revenue from stage $t+1$ until the end of the study horizon, seen from stage t at price level k. The approximations of the FBF are obtained iteratively by the dynamic programming algorithm. This calculation is made in a backward in time recursion based on the Bellman principle. When the problem of stage t and scenario k is solved, we obtain the optimal solutions for this stage and scenario and will be able to construct and approximation of the FBF for stage $t-1$ and so on, starting from the last stage of the horizon until the first one. When the backward recursion is complete, then a forward in time recursion is executed, optimizing the plant's operation in a Monte Carlo simulation, staring from the first stage until the last one, for all scenarios. This backward–forward procedure is performed as many times as needed to consider that the FBF approximations are satisfactory and the problem solution is good (for more details on the convergence of the algorithm see Chabar 2005). The dual variables associated to the constraints of reservoirs balance (A, B and of hours for next maintenance) are the opportunity costs of storage. The dual variable associated to the constraint of balance of reservoir A

$$(VA_{t+1} = VA_t + ARM_t - GToP_t + GTR_t - GDesc_t)\,,$$

for example, is the opportunity cost of storing the gas from the monthly ToP. The dual variable associated to the constraint of balance of each reservoir of remaining hours ($VH_{t+1}^{i,j} = VH_t^{i,j} \times (1 - X_t^{i,j}) + \overline{VH}^j \times (X_t^{i,j}) - \gamma \times EG_t^j$), reflects the opportunity cost of anticipating (by executing the maintenance before the number of hours reach the frequency of the cycle)/postponing (by not running or producing less with the unit) the execution of maintenance of cycle j in unit i.

3 Case Study

The application of the developed methodology will be illustrated here with a case study using data of the Brazilian system. The 80 spot price scenarios were obtained via a fundamentalist approach using the SDDP hydrothermal least-cost dispatch model (developed by PSR; Granville et al. 2003a, b; Pereira and Pinto 1984, 1985). The SDDP model was ran for a given configuration of the whole Brazilian system supply and demand and for a set of 80 samples of hydrological scenarios, generated by a stochastic inflow model starting on March 2004 and end on December 2008 in monthly time steps. A higher resolution of the operation "inside" the month is obtained by considering three load blocks (peak, off-peak, and an intermediate level).

The thermal plant considered has three closed-cycle gas-turbine (CCGT) units (with an efficiency of ~7,000 MMBTU/MWh) totaling 780 MW of installed capacity. The maintenance specifications are the ones in Table 1. It has been considered that this plant is, on average, unavailable 3% yearly due to unpredictable failures (EFOR), with 90% of the failures lasting 24 h and 10% of them lasting 360 h. The plant has an energy supply agreement of 725 MW at R$130/MWh and a FSA of 3.4 million m^3 of gas daily. In this contract, the percentages of monthly ToP, annual ToP, SoP, and DisCo margin are, respectively, 56, 70, 95, and 70%. The price of the commodity (gas) is R$3.5/MMBTU, the price of its transportation is R$4.5/MMBTU and the DisCo margin is R$0.5/MMBTU. The heat rate of the gas is 37,300 BTU/m^3 and the gas, if not consumed, may be stored by the fuel supplier for a maximum time of 7 years. It has been considered that the plant's fixed cost is R$3/kWmonth and that its variable cost (O&M) is R$4/MWh. We also assume that, at the beginning of the study horizon, no fuel was stored (reservoir A was empty).

In order to compare the flexible operation strategy (optimized by the model) and the usual operation policy (which establishes a monthly consumption of 70% of the gas contract – constant declaration of inflexibility without any storage of gas), we evaluate both strategies in both cases. The operation with the continuous consumption of 70% of the contracted gas does not consider the future benefit (FB) of the storage of the gas from ToP, since it does not consider opportunity costs. In this operation, all the components of the gas costs are faced as fixed costs. Nevertheless, the flexible operation calculated from the model, does consider this possible FB and takes it into account in the decision making of the operation. Moreover, with respect to the maintenance scheduling, the policy without FB (which does not consider opportunity costs) does not take into account the benefits that the anticipation of programmed interruptions brings. Consequently, all maintenance will take place when the units (individually) reach the number of hours of operation defined for each cycle. However this time may, unavoidably and undesirably, coincide with high price periods.

Two cases have been set up deriving from the general case: case1 – deterministic case with maintenance and case 2 – stochastic with maintenance. The deterministic case has been considered for an analysis purpose as they illustrate more straightforwardly the mechanisms of fuel storage and anticipation of maintenances.

3.1 Case 1: Deterministic with Maintenance

We consider, integrated to the management of the fuel supply contract, the optimization of maintenance scheduling. The operation strategy that considers the FB of fuel storage is more efficient than the one that fires all gas from ToP at each stage. This is achieved by turning off the generation units during the periods of low profitability (low prices) and storing fuel for future use. We see that through the comparison of the respective dispatches in Fig. 3a,b. Besides, Fig. 4a,b show that, during the periods of high prices when the plant wants to produce energy at "full throttle", the policy without FB, which does not provide any fuel storage. It can also be observed that, although the FB is considered, additional fuel (extra ToP) purchases take place. Nevertheless, these purchases are much lower than those that result from the policy without FB and happen since additional fuel is needed (during the period from September to December of 2006). That is, an operative policy with FB that results in the same financial outcome, but that does not buy additional gas during September to December of 2006 would be one that only anticipates the extra fuel consumption (like a "short blanket"). Moreover, if extra ToP purchases are unavoidable, it is desired that they happen the latest possible. This occurs because the strategy has a "wait-and-see" nature.

Concerning maintenance, the future signaling via the FB reallocates the scheduling, placing them in humid periods (low prices), while the policy without FB executes all maintenances always on the cycle limits, despite the spot price level at this moment. It can be observed comparing both operations in Fig. 5a,b. Besides, the operation that considers the FB, establishes a lower level of dispatch (the plant is only dispatched when the spot price is attractive and the amount of energy that is produced increases with the price). This becomes evident due to the fact that the first maintenance scheduled in the study horizon happens earlier than in the case without FB. This lower use of the plant's generation units, extending the time to the next maintenance, makes the policy with FB even more efficient than the conventional strategy. The FB function, in this case, provides an increase of approximately 7% in the plant's revenue, which jumps from R$845 millions (conventional case) to R$906 millions (case with FB).

3.2 Case 2: Stochastic with Maintenance

In reality, any isolated forecast of spot prices trajectory would be ignoring the inherent uncertainty of prices, which must be taken into consideration in the decision making process. The stochastic case is the one that better represents the reality. However, the consideration of the stochasticity of prices makes the problem more complex. In a stochastic environment, each decision does not just depend on the projected prices of a given series (as it happens in deterministic cases), but on all

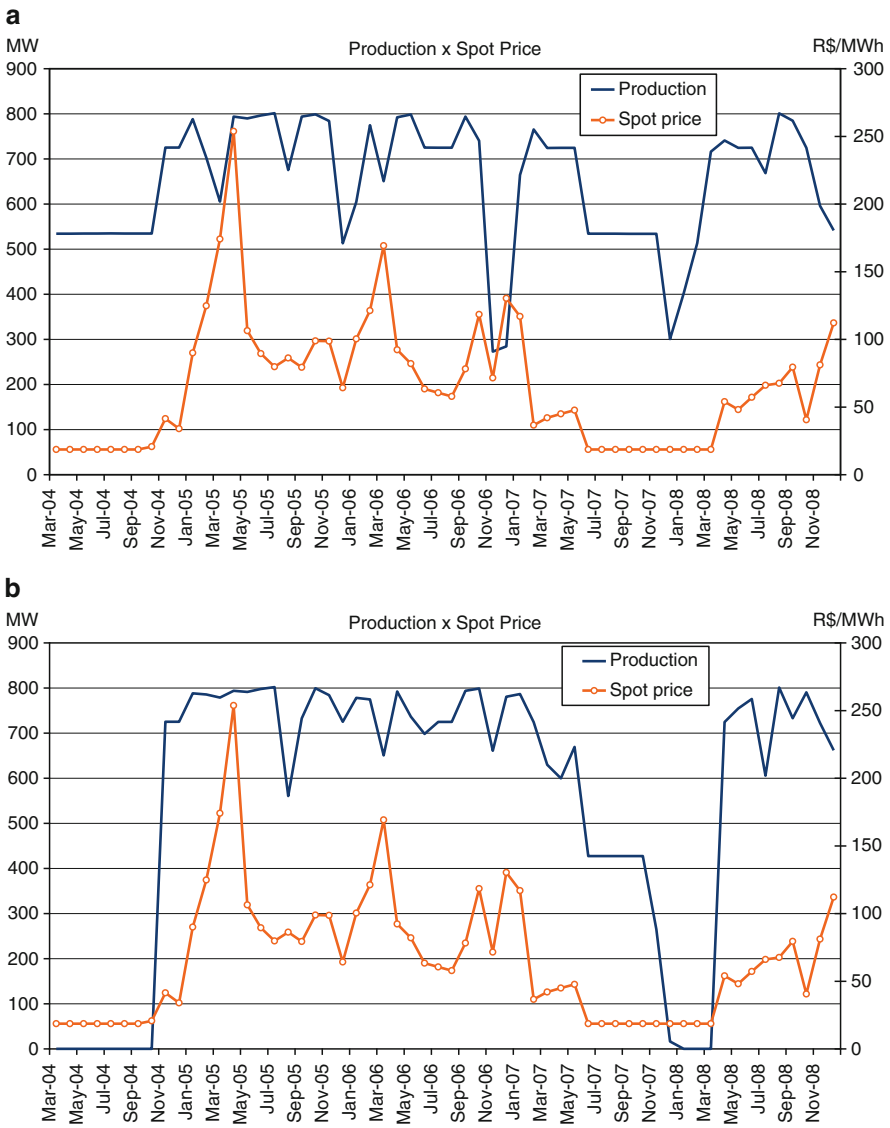


Fig. 3 (**a**) Optimization without future benefit function. (**b**) Optimization with future benefit function

possible future price evolution (considering the probability of occurrence of each evolution at each stage). The observed gain with the use of the FB function is now, on average, about 8%, increasing the expected value of the present value of the revenue from R$830 to R$893 millions.

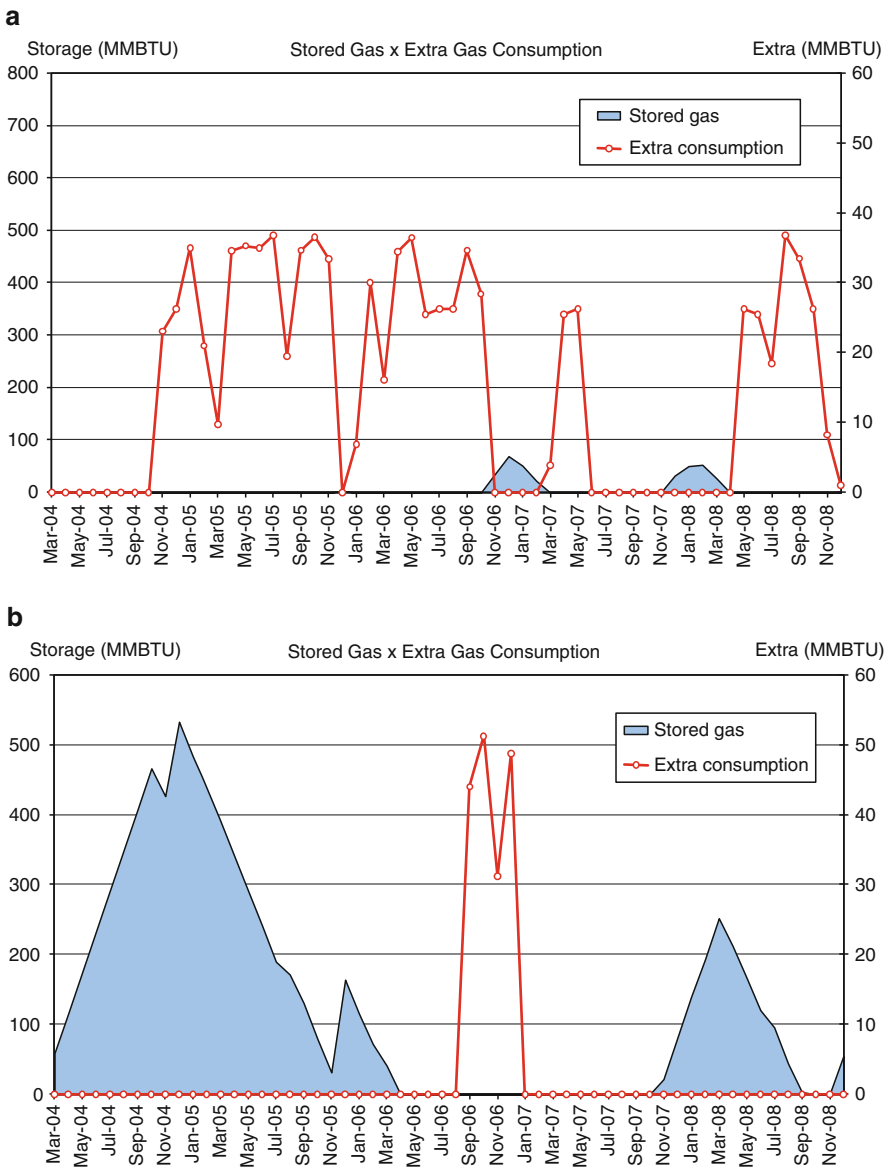


Fig. 4 (**a**) Optimization without future benefit function. (**b**) Optimization with future benefit function

Figure 6 shows the distribution of the present value of the plant's revenue for both policies. It can be observed that the curve that describes the distribution for the policy with FB is, for all scenarios, above the other curve that describes the policy without FB, stressing the gain that an efficient policy provides. This gain is

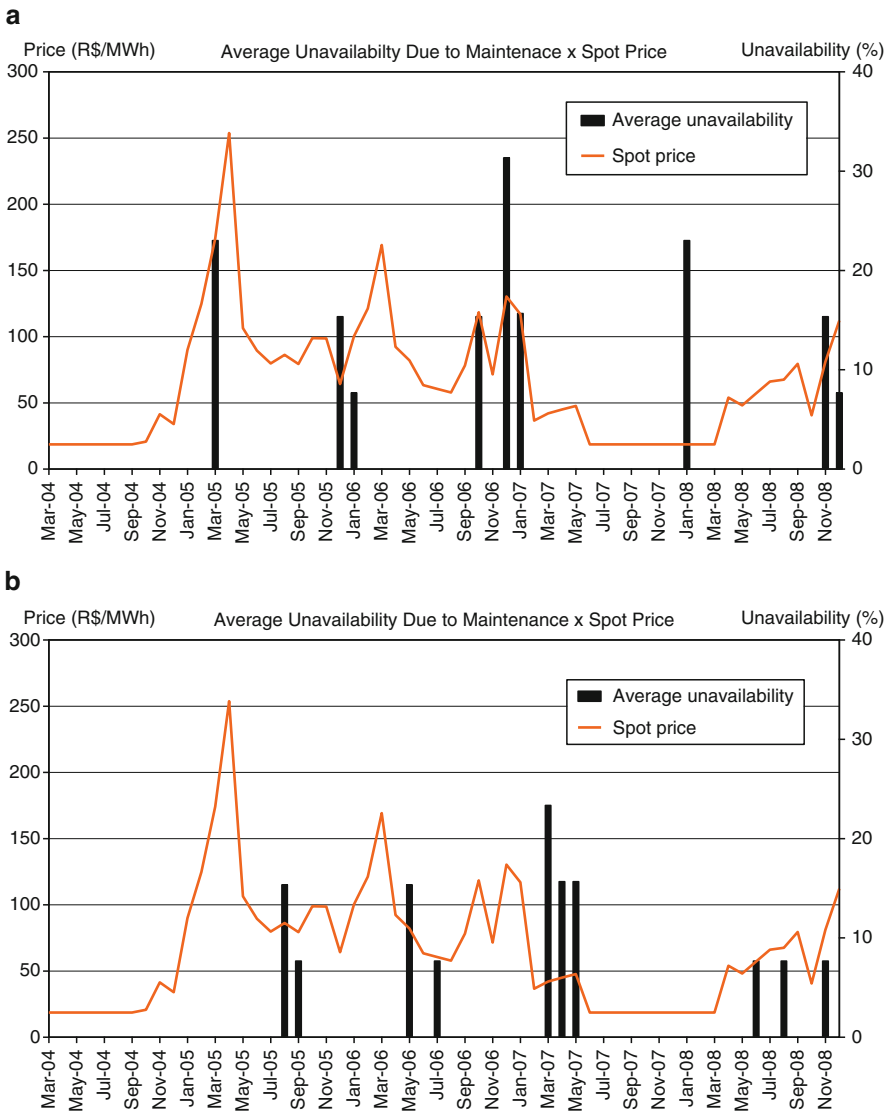


Fig. 5 (**a**) Optimization without future benefit function. (**b**) Optimization with future benefit function

especially high for the scenarios with the lowest spot prices, which are the ones that yield the lowest outcomes. On the other hand, for the highest values of revenues, which result from the scenarios that are conditioned to the highest level of prices, this difference becomes lower. For these situations, both policies (with and without FB) define similar dispatches, since the plant generates almost all the time if the spot price is high.

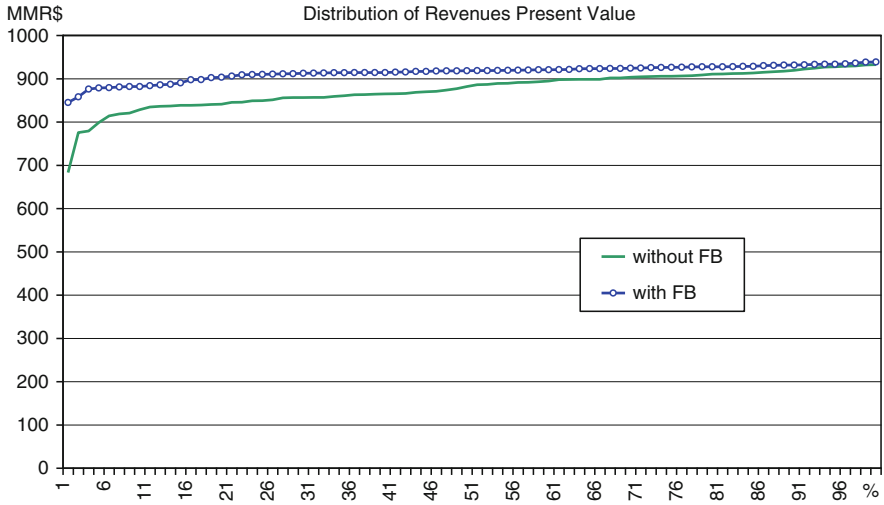


Fig. 6 Distribution of revenues present value

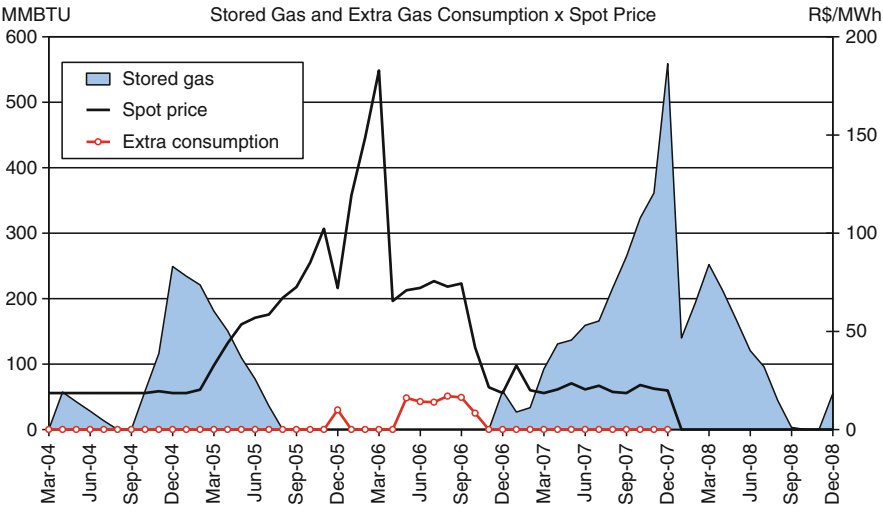


Fig. 7 Results for a given series (with future benefit function)

Figure 7 illustrates the decisions resulting from the policy with FB, of storage of gas and additional fuel consumption, for one of the 80 price scenarios that have been considered. If this scenario is compared to a deterministic case it could have been observed that the results in the stochastic case would have been different from the results in the deterministic case. The deterministic defined policy would have been the optimal policy for that specific scenario without considering the rest of the scenarios. In the stochastic case, however, several other possible spot price trajecto-

ries may occur and the decision to be taken is the one that is the best in terms of the expected value. This approach considers that it is unknown ex ante which of the 80 sample scenarios will occur.

4 European Application

In this section, we will describe the similarities between the above-analyzed Brazilian case and a European application. The differences between both markets are mainly associated to the development stage of the natural gas industry in those regions and consequently the market liquidity (Barroso et al. 2005). The section will be concluded with the adaptations and developments that have to be made to the model for this application.

4.1 Differences and Similarities

4.1.1 Contract Types

Gas contracts are both in Europe and Brazil of a take-or-pay type. Similar clauses for the monthly, weekly, and daily limits apply along with years-cycle make-up periods.

4.1.2 Markets

Market structure is the main difference that exists between the two cases. European market is liberalized with developed spot and financial markets. Products traded are liquid offering the opportunity to gas portfolio managers for a finer optimization. In the European market a gas plant is optimized taking into consideration the access to gas and power market for both buying and selling. The production of the plant and the gas contract are optimized comparing to the price level of the spark spread. Moreover, the European market has different price areas, which are interconnected and give a wider choice for power allocation.

4.1.3 Gas Contracts Portfolios

In contrast to the Brazilian market, the European market has a variety of long-term gas contracts of take-or-pay type. Each company can have a portfolio of these contracts in order to supply its power stations. This diversity demands for an optimization for the choice of contract used according to its clauses.

4.1.4 Corporate Behavior

Given the existence of the spot and financial markets, companies are able to design a strict risk policy. An elaborate risk policy results to dynamic hedging strategies where trading around the assets is essential. To structure dynamic strategies, a tool like the one presented above is needed.

4.1.5 Asset Portfolios

The asset portfolios in the European market are differently diversified, having a greater percentage of thermal plants and a much lower percentage of hydro. As mentioned in the second section, this structure gives another profile to price volatility making more difficult for the system to react to short-term than long-term shifts.

4.2 Algorithm of Future Adaptations and Developments

Further steps can be development in order to adapt the model for the European market. The future developments and adaptations are

- Gas market consideration for both buy and sell transactions,
- Financial market considerations for electricity and gas market using forwards and futures contracts for hedging purposes,
- Consideration of a gas contract portfolio for the supply of a portfolio of power stations,
- Consideration of gas transportation network constraints that may limit the use of sets of FSA in the gas contract portfolio to specific plants of the generation portfolio.

5 Conclusion

This work presented a computational model that determines the optimal dispatch strategy for a thermal plant with a gas contract. We have considered the specifications of the fuel supply agreement and its take-or-pay clauses, the opportunities of purchases and sales of energy in the spot market and the plant's detailed operational characteristics. In an integrated approach, the model determines also the optimal maintenance schedule for the plant's generation units. As the decisions in one stage have impact in the subsequent stages and as a function of the future trend of the short-term price, there is a time coupling in the decision making process. Thus, the problem is a multi-stage decision making under uncertainty. It has been shown that

a non-negligible gain can be obtained associated to the definition of an operative policy and maintenance scheduling strategy (instead of a fixed planning). This strategy seeks to manage, in an optimal way, the use of the fuel already bought (due to take-or-pay clauses) and determine the optimal maintenance schedule possible.

References

Barroso, L. A., Flach, B., Kelman, R., Binato, S., Bressane, J. M., & Pereira, M. (2005). *Integrated gas-electricity adequacy planning in Brazil: technical and economical aspects*. Proceedings of the IEEE General Meeting, San Francisco.

Chabar, R. M. (2005). *Otimização da operação e manutenção de usinas termelétricas sob incerteza em sistemas hidrotérmicos* MSc Dissertation, PUC-Rio, (jn Portuguese).

Flatabo, N., Haugstad, A., Mo, B., & Fosso, O. (1998). *Short and medium-term generation scheduling in the Norwegian hydro system under a competitive power market*. Proceedings of EPSOM Conference.

Granville, S., Kelman, R., Barroso, L. A., Chabar, R., Pereira, M. V., Lino, P., Xavier, P., & Capanema, I. (2003a). *Um sistema integrado para gerenciamento de riscos em mercados de energia elétrica*. XVII SNPTEE, Uberlândia, (in Portuguese).

Granville, S., Oliveira, G. C., Thomé, L. M., Campodónico, N., Latorre, M., Pereira, M. V., & Barroso, L. A. (2003b). *Stochastic optimization of transmission constrained and large scale hydrothermal systems in a competitive framework*. Proceedings of the IEEE General Meeting, Toronto. Available at http://www.psr-inc.com.

Gjelsvik, A., Belsnes, M., & Haugstad, A. (1999). *An algorithm for stochastic medium-term hydrothermal scheduling under spot price uncertainty*. Proceedings of 13th Power Systems Computation Conference.

Pereira, M. V. F., & Pinto, L. M. V. G. (1984). *Operation planning of large-scale hydrothermal systems*. Proceedings of the 8th PSCC, Helsinki, Finland.

Pereira, M. V. F., & Pinto, L. M. V. G. (1985). Stochastic otimization of multireservoir hydroelectric system – a decomposition approach. *Water Resource Research, 21*(6).

Pereira, M. V. F., Campodónico, N. M., & Kelman, R. (1998). *Long-term hydro scheduling based on stochastic models*. EPSOM'98, Zurich.

Wallace, S., & Fleten, S. E. (2003). *Stochastic programming models in energy. Stochastic Programming in the series Handbooks in Operations Research and Management Science* (Vol. 10, pp. 637–677).

Energy Portfolio Optimization for Electric Utilities: Case Study for Germany

Steffen Rebennack, Josef Kallrath, and Panos M. Pardalos

Abstract We discuss a portfolio optimization problem occurring in the energy market. Energy distributing public services have to decide how much of the requested energy demand has to be produced in their own power plant, and which complementary amount has to be bought from the spot market and from load following contracts. This problem is formulated as a mixed-integer linear programming problem and implemented in GAMS. The formulation is applied to real data of a German electricity distributor.

1 Introduction

We consider large German public services distributing energy in the order of magnitude of Düsseldorf, Hanover or Munich. On the one hand, the public services have to be large enough in order to utilize the optimization techniques discussed here; but on the other hand they have to be smaller than the supra-regional electric distributor, that is, RWE or E.ON.

The major difference of public services to supra-regional electric distributors is that public services usually do not sell excess energy in the energy market. They are price takers and their objective is to minimize the cost while meeting the demand for energy or electric power, resp.; in this paper we treat energy (physical unit: Wh or MWh) and electric power (physical unit: W or MW) as two different utilities which can be traded in the market.

S. Rebennack (✉) and P.M. Pardalos
Department of Industrial & Systems Engineering, Center for Applied Optimization, University of Florida, Weil Hall 303, P.O. Box 116595 Gainesville, FL 32611-6595, USA
e-mail: steffen@ufl.edu; pardalos@ufl.edu

J. Kallrath
Department of Astronomy, University of Florida, Weil Hall 303, P.O. Box 116595 Gainesville, FL 32611-6595, USA
e-mail: kallrath@astro.ufl.edu

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_14,
© Springer-Verlag Berlin Heidelberg 2010

The optimization model discussed in this article also does not apply to small public utility companies as they usually have one exclusive supplier of vendor, that is, RWE or E.ON. Therefore, they do not have a portfolio of sources of supply which can be optimized.

The considered electric distributor has several sources of supply in order to satisfy the demand for power of their customers. Among these possibilities are:

- The electric power generation in a single power plant operated autarkic by the electric distributor.
- The electric power generation in an external power plant. The operation of the plant is regulated by the carrier to a great extent.
- The purchase of energy in arbitrary quantities at any time from a business partner, known by name, with a bilateral treaty. This form of trading is called "Over The Counter." It stands in contrast to the anonymous stock jobbing.
- The purchase of standardized power products on the stock exchange, in the so-called *spot market*, abbreviated by SM. This is short-term trading.
- The purchase of power on the stock exchange in the forward market. This is long-term trading.
- The purchase of power in arbitrary quantities through so-called *load-following contracts* or LFCs.

The complete range of the opportunities can only be exploited in the mid-term; for instance, in an optimization over the whole year. In this article, we focus on the short-term portfolio optimization; that is, within one or two days. That is, we are given the operating conditions, including the mid-term decisions. The task is then to optimize the power plant operation and the purchase of energy in such a way that the total costs are minimized while satisfying the demand. The energy demand is given via a power forecast for the following day.

In this article, we develop a mixed-integer linear programming (MILP) formulation for the energy portfolio optimization problem allowing the following three sources of energy supply:

- The electric power generation in the own power plant,
- The purchase of standardized products from the spot market, and
- The purchase of power via the LFC with one supplier of vendor.

The mathematical programming formulation is implemented in the modelling language GAMS. The code has been added to the GAMS (2009) model library with the name `poutil.gms` (Portfolio Optimization for electric UTILities).

This electricity optimization problem falls in the scope of the *unit commitment problem* and *economic dispatch problem*. In contrast to the unit commitment problem, our model does not include any constraints on power transmission, reverse spinning or ramping. The economic dispatch problem differs from ours in the way that the different energy sources are only subject to capacity constraints whereas we have to deal with additional technical or production restrictions such as minimum idle time periods of the plant.

Dillon et al. (1978) provide a mixed-integer linear programming formulation of the unit commitment problem, also taking into account energy exchange contracts. The model by Carrion & Arroyo (2006) for thermal plants uses less binary variables than the model by Dillon et al. Our model assumes a discrete cost structure for the power plant in contrast to the quadratic one discussed by Carrion and Arroyo. Mixed-integer programming was also used by Hobbs et al. (2002) to solve the unit commitment problem. The optimal selling of energy in the electricity spot market is modelled as an MILP problem by Arroyo and Conejo (2000) and as a stochastic program by Philpott & Schultz (2006). In the literature, there are many specialized algorithms for solving the unit commitment problem (Wallace & Fleten (2003); Sheble & Fahd (1994); Padhy (2004); Sen & Kothari (1998); Baldick (1995)) and the economic dispatch problem (Madrigal & Quintana (2000); Chowdhury & Rahman (1990); Dhillon & Dhillon (2004)).

As we plan day-ahead, we assume that all data are reasonably well known. The day-ahead forecast is rather accurate but nevertheless subject to uncertainties. The forecast is derived from historical data, annual load profiles, weekday specifics tendencies, temperature profiles for the next days, and considers public holidays as well as special events such a soccer finals, formula I racings, etc. Smoothing and averaging over many influence factors lead to a rather stable forecast. The remaining uncertainties are of the order of a few percent and may lead to minor changes; they are mostly covered by LFC costs. The prices for the purchased energy are given through contracts and the spot market. Furthermore, we assume to have a quite accurate power forecast for the planning horizon. However, when such data are not reliable or when looking at longer planning horizons, a stochastic model would be preferable against a deterministic one; taking into account, for instance, the stochastic spot prices and/or stochastic demand. Such models and algorithms are discussed, for instance, by Takriti et al. (1996, 2000) and by Shiina & Birge (2004). Including hydro, wind or solar as an energy source into the model leads also to stochastic components; c.f. Nowak & Römisch (2000); Gröwe-Kuska & Römisch (2005); Brand et al. (2004).

A simple unit commitment model code is available in the LINDO Systems (2003) library model `unitcom1.lg4`.

We start with the description of the problem in Sect. 2. The mathematical formulations are discussed in detail in Sect. 3, including the special cost structure of the different energy sources and the constraints associated with the power plant operation. In Sect. 4, we discuss some limitations of the model and provide possible modifications of the formulation. Computational results for the implemented model in GAMS are given in Sect. 5. Conclusions of this article are provided in Sect. 6.

Throughout the article, we introduce several sets, variables and input data given. We denote all variables with small letters and input data as capital ones. In the Appendices, all sets (App. A), variables (App. B), constraints (App. C), input data and parameters (App. D) used in the mathematical model are summarized along with their synonyms in the GAMS (2009) model `poutil.gms`.

2 Description of the Problem

In this section, we discuss the short-term optimization problem for the day-ahead planning of the energy portfolio.

In general, the power curve of one day is given by the continuous function

$$P(t), \qquad 0 \leq t \leq 24,$$

given in MW. We brake the power process into quarters of an hour. The use of quarter-hour values as general time frame is a common standard in worldwide energy economics; furthermore it is based on several directives, as, for instance, in Germany the MeteringCode (VDN (2006)), in Austria the statistical regulation Österreichische Elektrizitätsstatistikverordnung (2007); as a practical example one can find the published maximum load values of Stadtwerke Saarlouis GmbH (2003) in quarter-hours. Furthermore, in the energy industry, the continuous process of the produced and provided power is treated as fixed within a quarter-hour basis. With this convention, we can approximate the power curve through a step function. Let $\mathcal{T}$ be the set of quarter-hour time slices per day; that is, $\mathcal{T} := \{1, \ldots, N^{\mathrm{T}} = 96\}$. We assume that we are given the forecast of electric power for all 96 quarter-hour time intervals per day

$$P_t, \qquad t = 1, \ldots, N^{\mathrm{T}},$$

measured in MW. In order to meet the demand, the utility company disposes of three sources of supply,

- A power plant (PP) with given capacity.
- The opportunity to buy power from the spot market at the energy bourse in the form of standardized products.
- A load-following contract with one supplier of vendor. The amount of energy is assumed to be unlimited.

The total cost for the fulfillment of the demand is then given by the sum of the power plant operation cost, the cost for the purchase of power from the spot market and the cost for the purchase of power from the open supply contract.

The structure of the cost components and the constraints involved are discussed in the following sections.

2.1 Power Plant Usage

We assume that we are given a natural gas power plant. The reasons are that they are quite common in Germany (23% of primary energy supply in 2004; European Commission (2007)) and that they can be operated very flexibly. This implies that we do not have to consider restrictions which last for more then one day.

The costs of the power generation in the own power plant consist in principle of the fix costs per day and the variable costs per MWh generated. To simplify matters, the variable costs of the power generation are assumed to be constant. This disregards that operational costs depend on the actual degree of efficiency and that operating a power plant beside the point of optimum causes variable costs to increase; see Sect. 4.2 for further details.

Let us now discuss the constraints associated with the power plant usage. The power plant has a maximal power of $P^{\mathrm{PP}}_{\max}$, measured in MW. During normal operation, the power plant should not be operated with less than 40% of its maximal power. This is not a technical restriction or a generally accepted convention, but a useful approach to avoid an obvious contradiction to the assumption of constant variable costs.

Let p^{PP}_t be the amount of power in MW of the power plant at time period t. Then we get

$$p^{\mathrm{PP}}_t \geq 0.4 P^{\mathrm{PP}}_{\max}, \qquad \forall t, \tag{1}$$

in case the power plant is used; otherwise we have $p^{\mathrm{PP}}_t = 0$, obviously.

For technical reasons, the power of the plant is not a continuous variable but fixed in steps of 10% of the maximal power. A restriction to 10% steps while running a power plant is obviously deliberate but one should remember that an operator would never choose an infinite continuum of steps but only a small number of usual operating points. These so-called *partial load operation points* are ordinarily determined by technical attributes of the power plant and are supposed to be given. Whether these in our model are defined as equidistant steps or as a set of given figures does not matter. However, it is important to define them as a small set of discrete numbers to approach reality.

Define stage 1 as the idle stage of the plant and stages 2, 3, ..., 8 as the stages corresponding to the power level of 40% $P^{\mathrm{PP}}_{\max}$, 50% $P^{\mathrm{PP}}_{\max}$, ..., 100% $P^{\mathrm{PP}}_{\max}$. The stages and the corresponding power level with respect to the maximal power level $P^{\mathrm{PP}}_{\max}$ are illustrated in Fig. 1. This allows us to substitute (1) by

$$p^{\mathrm{PP}}_t = 0.1(\alpha_s + 2) P^{\mathrm{PP}}_{\max}, \qquad \forall t \tag{2}$$

with $\alpha_s \in \{2, 3, 4, 5, 6, 7, 8\}$.

In order to avoid permanent changes of the power level, we require any power stage to continue for at least D^{PP}_{act} quarter-hours, with a typical value of $D^{PP}_{\mathrm{act}} = 8$. A constant operation over a period of $D^{\min}_{\mathrm{act}}$ quarter-hours is a deliberate simplification of the model as well; but it covers the experience that it could be considered as ineffective to change the operation mode of an engine permanently. The change itself causes loss of energy through *start-up* and *shut-down losses* (Wood & Wollenberg (1996)), which we do not want to take into consideration here. This restriction on the changes of the power plant can be formulated as

$$p^{\mathrm{PP}}_j = p^{\mathrm{PP}}_{j+1} = \ldots = p^{\mathrm{PP}}_{j+k}, \quad \text{with } k \geq 7, \tag{3}$$

where j is a time interval containing a shift of the power level.

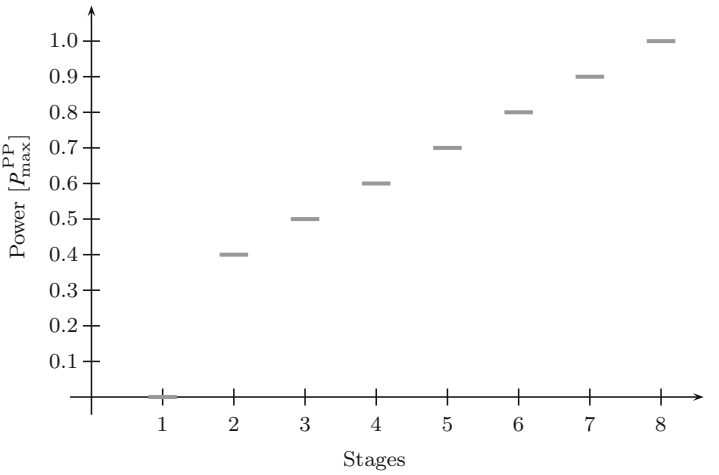


Fig. 1 Stages of the power plant vs. fraction of maximal power level

To avoid a complete shut-down of the power plant for only a short time period, any idle period has to last for at least 4 h:

$$p_j^{PP} = p_{j+1}^{PP} = \ldots = p_{j+m}^{PP}, \quad \text{with } m \geq 15, \tag{4}$$

where j is a time interval containing an idle time.

We relax this condition for the end of a day. The idle times can then be shorter, as some part of this time can be transformed to the next day or may be coming from the previous one. These boundary conditions show the drawback of looking at each day separately. In reality, every day has some pre-history, providing the boundary conditions.

2.2 Energy Purchase from the Spot Market

The European Energy Exchange (EEX) in Leipzig provides the spot market as an opportunity to trade energy. This means that we can buy standardized products in short term. We consider here the so-called *base load* and *peak load* contracts which belong to the continuous trading of EEX[1]; Madlener & Kaufmann (2002). They are traded at one day and delivered at the next day; EEX (2007). Special cases occurring, for instance, on weekends are not considered here; those are the weekend-base load contracts[2].

[1] We do not consider selling in the auction market in our model.

[2] Weekend-base load contracts specify the delivery for 48 h, starting at Saturday 0:00 a.m. and ending on Sunday 12:00 p.m.; peak load contracts for the weekends are not offered.

Each base load contract specifies the delivery of a constant power of 1 MW from 0:00 a.m. to 12:00 p.m. at the following day after the completion of the contract.

Each peak load contract specifies the delivery of a constant power of 1 MW from 8:00 a.m. to 8:00 p.m. at the following day after the completion of the contract.

The provider and customer remain anonymous for these contracts. The commercial clearing and settlement is handled by the EEX, while the technical delivery is done through the power grid operators in Germany. Currently, the power grid in Germany is not uniform nationwide. There are four transmission network operators: E.ON, Vattenfall, RWE Transportnetz Strom and EnBW.

We get from the conventions above that the contribution to the energy portfolio from the spot market, e_t^{SM}, is given though the number α of base load and the number β of peak load contracts bought, while respecting the above time intervals for energy delivered.

The cost for the energy from the spot market is calculated via the total delivered energy amount in MWh.

2.3 Energy Purchase from the Load-Following Contract

The LFC can be seen as a compensation for the vacancy of the previously discussed sources of energy supply; Heuck & Dettmann (2005). An energy load can be covered only partially by the standardized products from the spot market and the relatively inflexible power plant operation. However, the utility company is committed to meet the power demand of its customers. Therefore, the vacancy has to be closed by a flexible instrument. Obviously, the flexibility of this instrument makes the energy purchase from the LFC to the most expensive source of the three discussed in the paper as it transfers all risk from the customer to the seller of the contract. The LFCs are also called *full requirements contracts*.

The costs for the LFC are determined via the typical two-component supply-contracts; Erdmann & Zweifel (2007). That is, the delivered power, or more precisely, the power level peak, as well as the delivered energy amount, are considered. In other words, it is the sum of the so-called *power rate* [€/MW] and the *energy rate* [€/MWh].

The power rate $C_{\mathrm{PR}}^{\mathrm{LFC}}$ of the LFC is based on the highest drain of power (quarter-hour value) within a year $p_{\max}^{\mathrm{LFC}}$. To avoid random anomalies up to a certain amount, one usually applies the arithmetic mean of the two – in some contracts also three – highest monthly peaks as the rated value of the calculation of the power rate. We get for the cost of the power rate

$$C_{\mathrm{PR}}^{\mathrm{LFC}} = C_{\mathrm{PR,year}}^{\mathrm{LFC}} \cdot p_{\max}^{\mathrm{LFC}}, \tag{5}$$

where $C_{\mathrm{PR,year}}^{\mathrm{LFC}}$ is the cost coefficient per MW of the power rate on an annual basis.

For the demand rate contracts considered in this article, usually there are defined annually quantity zones with different prices. Let Z_1 and Z_2 be the borders of the

quantity zones given in MWh and let P_1^{LFC}, P_2^{LFC} and P_3^{LFC} be the prices in € per MWh in these zones. We denote by $e_{\text{year}}^{\text{LFC}}$ the delivered energy amount annually. Then, the prices in € per MWh are given by

$$\left\{ \begin{array}{l} P_1^{\text{LFC}}, \text{ if } 0 \leq e_{\text{year}}^{\text{LFC}} \leq Z_1 \\ P_2^{\text{LFC}}, \text{ if } Z_1 < e_{\text{year}}^{\text{LFC}} \leq Z_2 \\ P_3^{\text{LFC}}, \text{ if } Z_2 < e_{\text{year}}^{\text{LFC}} \end{array} \right\}.$$

Recognize that the price P_1^{LFC} is paid for the amount of energy in zone 1, where price P_2^{LFC} is only paid for the amount of energy within zone 2, exceeding the quantity in zone 1.

The quantity price $P_{\text{year}}^{\text{LFC}}$, or total variable cost per year associated with the LFC, can then be stated as

$$P_{\text{year}}^{\text{LFC}} = \left\{ \begin{array}{ll} P_1^{\text{LFC}} \cdot e_{\text{year}}^{\text{LFC}}, & \text{if } 0 \leq e_{\text{year}}^{\text{LFC}} \leq Z_1 \\ P_1^{\text{LFC}} \cdot Z_1 + P_2^{\text{LFC}} \left(e_{\text{year}}^{\text{LFC}} - Z_1 \right), & \text{if } Z_1 < e_{\text{year}}^{\text{LFC}} \leq Z_2 \\ P_1^{\text{LFC}} \cdot Z_1 + P_2^{\text{LFC}} \left(Z_2 - Z_1 \right) + P_3^{\text{LFC}} \left(e_{\text{year}}^{\text{LFC}} - Z_2 \right), & \text{if } Z_2 < e_{\text{year}}^{\text{LFC}} \end{array} \right..$$

The resulting piece-wise linear price curve is shown in Fig. 2.

This price system is adjusted annually. When using it on a daily basis, it leads to the following effect. At the beginning of the year, we are always in zone 1, growing steadily into zone 2 and resulting finally in zone 3 at a particular point of time. With this interpretation of the model, the effective current price depends on the relative position of the day within the year. This leads to difficulties for short-term modeling. To overcome this problem, we introduce a daily-based model in Sect. 3.1.

The amount of energy from LFC is, in principle, unlimited and can vary in each of the quarter-hour time periods without restrictions. Hence, no additional constraints for the energy purchase from the LFC are needed.

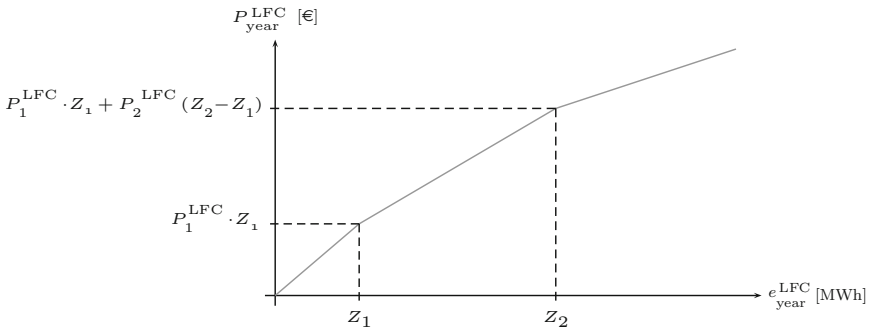


Fig. 2 Piece-wise linear price curve for the load-following contract (on an annual basis)

3 Mathematical Formulation

In this section, we formulate the described problem above as a MILP problem. Our task is to minimize the total cost while meeting the demand forecast for each quarter-hour time interval and the constraints associated with the power plant usage.

3.1 Objective Function

The total cost c^{tot} for the fulfillment of demand for the particular day consists of the cost for the power plant operation, c^{PP}, the cost for the purchase of power from the spot market, c^{SM}, and the cost for the purchase of power from the load-following contract, c^{LFC}. Hence, we get for the total cost

$$c^{\text{tot}} = c^{\text{PP}} + c^{\text{SM}} + c^{\text{LFC}}. \tag{6}$$

Let us now discuss the three cost components in detail.

3.1.1 Cost for the Power Generation in the Own Power Plant

The cost associated with the power plant is given by the sum of the fixed cost $C^{\text{PP}}_{\text{fix}}$ and the variable cost $C^{\text{PP}}_{\text{var}}$ per MWh. Recognize that the variable cost represents the cost for the produced energy and the fixed cost includs the electric power cost; that is, the power capacity of the plant influences the construction cost of the plant which are included in the fixed cost $C^{\text{PP}}_{\text{fix}}$. We can then write the total cost in € as

$$c^{\text{PP}} = C^{\text{PP}}_{\text{fix}} + C^{\text{PP}}_{\text{var}} \cdot e^{\text{PP}}, \tag{7}$$

where e^{PP} is the total energy withdrawn from the power plant. If we denote by p^{PP}_t the electric power in MW of the power plant during time slice t, then we get

$$e^{\text{PP}} = \sum_{t=1}^{N^{\text{T}}} \frac{1}{4} p^{\text{PP}}_t. \tag{8}$$

3.1.2 Cost for the Purchase of Energy from the Spot Market

As introduced in Sect. 2.2, let α be the number of base load and β be the number of peak load contracts. The electric power purchased per time interval t (quarter-hour) is then given by

$$p^{\text{SM}}_t = \alpha + I^{\text{PL}}_t \cdot \beta, \tag{9}$$

with the usage of step function I_t^{PL} for the peak load contracts. From the description of Sect. 2.2, they are active within 48 quarter-hour intervals, for respectively 12 h

$$I_t^{\mathrm{PL}} = \begin{cases} 0, & t = 1, \ldots, 32 \quad \text{and} \quad t = 81, \ldots, 96 \\ 1, & t = 33, \ldots, 80 \end{cases}. \tag{10}$$

The payment has to be made over the total energy amount in MWh delivered, resulting in

$$e^{\mathrm{SM}} = \sum_{t=1}^{N^{\mathrm{T}}} \frac{1}{4} p_t^{\mathrm{SM}} = \sum_{t=1}^{N^{\mathrm{T}}} \frac{1}{4} \left(\alpha + I_t^{\mathrm{PL}} \cdot \beta\right) = 24 \cdot \alpha + 12 \cdot \beta. \tag{11}$$

Finally, the cost for the purchase of energy from the spot market during the day are determined by the bourse. They are C^{BL} € per MWh for the products base load and C^{PL} € per MWh for peak load, respectively. Finally, this yields to the cost

$$c^{\mathrm{SM}} = \sum_{t=1}^{N^{\mathrm{T}}} \frac{1}{4} \left(C^{\mathrm{BL}} \cdot \alpha + C^{\mathrm{PL}} \cdot I_t^{\mathrm{PL}} \cdot \beta\right) = 24 \cdot C^{\mathrm{BL}} \cdot \alpha + 12 \cdot C^{\mathrm{PL}} \cdot \beta, \tag{12}$$

associated with the purchase of energy from the spot market. As the electric power for the base load and peak load contracts is constant, there is no additional cost for the electric power associated with the base load and peak load contracts.

3.1.3 Cost for the Energy Purchase from the Load-Following Contract

In Sect. 2.3, we saw that the price of the LFC is given as the sum of the power rate and the variable cost per MWh purchased, the energy rate.

The power rate $C_{\mathrm{PR}}^{\mathrm{LFC}}$ is given through formula (5), which depends on the maximum yearly power level $p_{\max}^{\mathrm{LFC}}$ with respect to quarter-hours. Notice that optimization could lead to the scenario that for a short time period high power is drained, which contributes only very little energy but results in high energy peaks implying a high power rate. In order to avoid such situations, we introduce an electric power reference level, $P_{\mathrm{ref}}^{\mathrm{LFC}}$, which is not allowed to be exceeded by the electric power purchased from the LFC. This reference level could either be the highest measured value so far, a corresponding last year value, an arbitrary limit which is not allowed to be exceeded, or a reference level determined by a mid-term/long-term optimization model. Hence, we want to satisfy the following constraint

$$p_t^{\mathrm{LFC}} \leq P_{\mathrm{ref}}^{\mathrm{LFC}}, \qquad \forall t, \tag{13}$$

with p_t^{LFC} being the electric power from the LFC for time slice t. This hard constraint on p_t^{LFC} allows us to substitute $p_{\max}^{\mathrm{LFC}}$ in formula (5) by $P_{\mathrm{ref}}^{\mathrm{LFC}}$. Hence, the power rate reduces to fixed cost on an annual basis. As our model is a short-term optimization model, these costs are not relevant. Therefore, the cost for the purchase from the LFC is given by the energy rate $c_{\mathrm{ER}}^{\mathrm{LFC}}$, which is variable cost per MWh, as

$$c^{\mathrm{LFC}} = c_{\mathrm{ER}}^{\mathrm{LFC}}. \tag{14}$$

Now, consider the special zone prices of the LFC described in Sect. 2.3. As already mentioned, the annually based price system is improper for our optimization model. To overcome these difficulties, we split the zones into daily quantities and simulate daily zones. Instead of using Z_1 and Z_2, the zonal borders Z_1^{d} and Z_2^{d} are utilized with

$$Z_1^{\mathrm{d}} = Z_1/365, \quad Z_2^{\mathrm{d}} = Z_2/365. \tag{15}$$

With e^{LFC} as the daily delivery quantity from the load-following contract

$$e^{\mathrm{LFC}} := \sum_{t=1}^{N^{\mathrm{T}}} \frac{1}{4} p_t^{\mathrm{LFC}}, \tag{16}$$

we have that the quantity price of one day is given by

$$c^{\mathrm{LFC}} = \begin{cases} P_1^{\mathrm{LFC}} \cdot e^{\mathrm{LFC}}, & \text{if } 0 \le e^{\mathrm{LFC}} \le Z_1^{\mathrm{d}} \\ P_1^{\mathrm{LFC}} \cdot Z_1^{\mathrm{d}} + P_2^{\mathrm{LFC}} \left(e^{\mathrm{LFC}} - Z_1^{\mathrm{d}}\right), & \text{if } Z_1^{\mathrm{d}} < e^{\mathrm{LFC}} \le Z_2^{\mathrm{d}} \\ P_1^{\mathrm{LFC}} \cdot Z_1^{\mathrm{d}} + P_2^{\mathrm{LFC}} \left(Z_2^{\mathrm{d}} - Z_1^{\mathrm{d}}\right) + P_3^{\mathrm{LFC}} \left(e^{\mathrm{LFC}} - Z_2^{\mathrm{d}}\right), & \text{if } Z_2^{\mathrm{d}} < e^{\mathrm{LFC}} \end{cases}.$$

In order to keep the model generic, we assume to have N^{B} different zones; where $b \in \mathcal{B}$ is one of the zones; that is, $b \in \mathcal{B} := \{1, \ldots, N^{\mathrm{B}}\}$. In our case, we have $N^{\mathrm{B}} = 3$. To identify the appropriate price segments, we use the binary variables μ_b. These variables indicate in which interval the daily purchased amount of energy lies, that is

$$\mu_b := \begin{cases} 1, & \text{if } Z_{b-1}^{\mathrm{d}} \le e^{\mathrm{LFC}} < Z_b^{\mathrm{d}} \\ 0, & \text{otherwise} \end{cases}, \qquad b = 1, \ldots, N^{\mathrm{B}}, \tag{17}$$

where we define for notational convenience $Z_0^{\mathrm{d}} = 0$ and $Z_{N^{\mathrm{B}}}^{\mathrm{d}}$ as a number large enough. Let variable e_b^{LFC} be the contribution to e^{LFC} in segment b. Then we get that the equalities

$$\sum_{b=1}^{N^{\mathrm{B}}} \mu_b = 1 \tag{18}$$

and

$$e^{\mathrm{LFC}} = \sum_{b=1}^{N^{\mathrm{B}}} \left(Z_{b-1}^{\mathrm{d}} \mu_b + e_b^{\mathrm{LFC}} \right), \tag{19}$$

as well as the inequalities

$$e_b^{\mathrm{LFC}} \leq \left(Z_b^{\mathrm{d}} - Z_{b-1}^{\mathrm{d}} \right) \mu_b, \qquad b = 1, \ldots, N^{\mathrm{B}} \tag{20}$$

connect variables e_b^{LFC} and μ_b to the energy e^{LFC} purchased from the LFC. Hence, we obtain the energy rate of the LFC

$$c_{\mathrm{ER}}^{\mathrm{LFC}} = \sum_{b=1}^{N^{\mathrm{B}}} \left(C_b^{\mathrm{LFC}} \cdot \mu_b + P_b^{\mathrm{LFC}} \cdot e_b^{\mathrm{LFC}} \right), \tag{21}$$

where C_b^{LFC} are the accumulated cost up to segment b, that is,

$$C_b^{\mathrm{LFC}} = \begin{cases} 0, & \text{if } b = 1 \\ P_1^{\mathrm{LFC}} \cdot Z_1^{\mathrm{d}}, & \text{if } b = 2 \\ C_{b-1}^{\mathrm{LFC}} + P_{b-1}^{\mathrm{LFC}} \left(Z_{b-1}^{\mathrm{d}} - Z_{b-2}^{\mathrm{d}} \right), & \text{if } b = 3, \ldots, N^{\mathrm{B}} \end{cases} \tag{22}$$

The breaking down of the zone prices on a daily basis is a trick to present the special price structure of the LFC. In practice, one could use the data of the previous years to estimate the cost of the LCF for each day. However, such a method requires a huge amount of experience in order to adjust the price in a meaningful way and it has to be seen in practice if it would outperform the special modelling of the zone prices discussed above.

The set of variables $\mu_1, \ldots, \mu_{N^{\mathrm{B}}}$ form a so-called *Special Order Set of type 1* (SOS-1), as only one variable of the set can have a non-zero value. The SOS-1 was introduced by Beale & Tomlin (1969). Description of SOS-1 in the context of integer programming can be found, for instance, in Kallrath & Wilson (1997, Chapter 6.7).

3.2 Demand and Power Plant Constraints

Let us now discuss the demand constraints and the constraints for the power plant operation.

3.2.1 Power Demand Constraints

Clearly, we have to meet the electric power demand for each quarter-hour. That gives us

$$p_t^{\mathrm{PP}} + p_t^{\mathrm{SM}} + p_t^{\mathrm{LFC}} = P_t, \qquad t = 1, \ldots, N^{\mathrm{T}}. \tag{23}$$

Recognize that the power demand has to be met exactly. The reason is that (at least a large amount of) energy cannot be stored.

3.2.2 Power Plant Constraints

We have to discuss the modelling of the restricted operation of the power plant. Therefore, we introduce the binary variables

$$\delta_{mt} := \begin{cases} 1, \text{ if the power plant is at time } t \text{ at stage } m \\ 0, \text{ otherwise} \end{cases} \tag{24}$$

to model the stages, $m \in \mathcal{M} := \{1, \ldots, N^{\mathrm{M}} = 8\}$, of the plant. Stage $m = 1$ corresponds to the idle state of the power plant. Values $m = 2, \ldots, N^{\mathrm{M}} = 8$ refer to the capacity utilizations 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1, respectively. The plant is in exactly one of those stages at any time, that is

$$\sum_{m=1}^{N^{\mathrm{M}}} \delta_{mt} = 1, \qquad \forall t. \tag{25}$$

The utilized power can then be calculated according to the following formula

$$p_t^{\mathrm{PP}} = P_{\max}^{\mathrm{PP}} \cdot \sum_{m=2}^{N^{\mathrm{M}}} \frac{1}{10} (m+2)\, \delta_{mt}, \qquad \forall t, \tag{26}$$

where $P_{\max}^{\mathrm{PP}}$ is the capacity of the power plant in MW. Note that this is the counter part of (2) with binary variables, but also holds true when the plant is in the idle stage 1.

In (3), we formulated the requirement that any power stage has to be continued for at least 2 h. This constraint is called *minimum up time constraint*. For this purpose, the binary variables χ_t^{S} keep track, if there is a change in the power plant level in time slice t

$$\chi_t^{\mathrm{S}} \geq \delta_{mt} - \delta_{mt-1}, \qquad \forall m, \quad t = 2, \ldots, N^{\mathrm{T}}, \tag{27}$$

and

$$\chi_t^{\mathrm{S}} \geq \delta_{mt-1} - \delta_{mt}, \qquad \forall m, \quad t = 2, \ldots, N^{\mathrm{T}}. \tag{28}$$

Inequalities (27) and (28) ensure that variable χ_t^{S} has value 1, if there is a change in the stage of the plant; however, χ_t^{S} can also have value 1, if there was no change in the stage. It is only important that it is now possible to formulate the condition

$$\chi_t^{\mathrm{S}} + \chi_{t+1}^{\mathrm{S}} + \chi_{t+2}^{\mathrm{S}} + \chi_{t+3}^{\mathrm{S}} + \chi_{t+4}^{\mathrm{S}} + \chi_{t+5}^{\mathrm{S}} + \chi_{t+5}^{\mathrm{S}} + \chi_{t+6}^{\mathrm{S}} + \chi_{t+7}^{\mathrm{S}} \leq 1, \quad t = 1, \ldots, N^{\mathrm{T}} - 7$$

or generally

$$\sum_{k=1}^{D_{\mathrm{act}}^{\mathrm{PP}}} \chi_{t+k-1}^{\mathrm{S}} \leq 1, \qquad t = 1, \ldots, N^{\mathrm{T}} - \left(D_{\mathrm{act}}^{\mathrm{PP}} - 1\right), \tag{29}$$

ensuring that within any 2 h, or $D_{\mathrm{act}}^{\mathrm{PP}} = 8$ time intervals, at most one stage change takes place.

In addition to the restrictions above, we discussed in Sect. 2.1 also the requirement for any idle period to be at least 4 h. This condition is called *minimum idle time requirement* or *minimum down time requirement*. Let us introduce the binary variable χ_t^{I}, indicating if the power plant has been started, that is, if it left the idle state in time slice t. We get the following inequalities

$$\chi_t^{\mathrm{I}} \geq \delta_{1t-1} - \delta_{1t}, \qquad t = 2, \ldots, N^{\mathrm{T}}. \tag{30}$$

The condition for the idle period given in (4) can then be modelled as

$$\sum_{k=1}^{D_{\mathrm{idl}}^{\mathrm{PP}}} \chi_{t+k-1}^{\mathrm{I}} \leq 1, \qquad t = 1, \ldots, N^{\mathrm{T}} - \left(D_{\mathrm{idl}}^{\mathrm{PP}} - 1\right) \tag{31}$$

with $D_{\mathrm{idl}}^{\mathrm{PP}} = 16$, or 4 h respectively. Constraint (31) can be interpreted in the way that the power plant is not allowed to leave the idle state more than once within any $D_{\mathrm{idl}}^{\mathrm{PP}}$ time slices.

As already mentioned in Sect. 2.1, we relax the condition of the minimum up and idle time for the beginning and the end of the planning horizon. However, for $t = 1$, $t = N^{\mathrm{T}} - \left(D_{\mathrm{act}}^{\mathrm{PP}} - 1\right)$, we have that the stage of the power plant is allowed to change only once in the first and last $D_{\mathrm{act}}^{\mathrm{PP}}$ time slices.

The variables χ_t^{S} are initially binary variables indicating a change of the stage of the power plant. However, we can relax these variables to be non-negative continuous. The reason is that constraints (27), (28) and (29) force the variables χ_t^{S} to be binary in the case that the minimum up time condition is tight, as the right hand side of constraints (27), (28) and (30) can only take the values 0 and 1. Recognize that this does not mean that the left hand side of constraints (29) being equal to 1 implies that the variables χ_t^{S} are binary. From the modeling point of view, it is therefore equivalent to use a binary or a non-negative continuous domain for variables χ_t^{S}. However, computationally, there is a difference[3]. The reason is that most

[3] For the real data of Stadtwerke Saarlouis, the running time of the continuous model compared to the binary model was less than 40%, it needed 45% of the iterations and 60% of the branching nodes.

Branch & Bound and Branch & Cut algorithms use LP domain relaxations, treating binary variables as continuous; c.f. Wolsey & Nemhauser (1999) and Atamtürk & Savelsbergh (2005). The branching process ensures then that those continuous variables are forced to be integral. In case of variable χ_t^{S}, we do not want the solver to branch on those, as their integrality is already applied by the binary variables δ_{mt}. However, if we can "forbid" the solver to branch on those variables (in GAMS this is accomplished by setting the priorities to `+inf`), then these two approaches of modelling the domain are also computationally equivalent[4]. The same concept holds also true for the variables χ_t^{I}.

This idea of avoiding to branch on variables χ_t^{S} and χ_t^{I} can be realized in the modelling language GAMS by defining branching priorities for these variables; c.f. Rosenthal (1997, 2008) and Bruce et al. (2009). The default branching priority for integral variables in GAMS is value 1. The higher the value, the lower is the priority to branch on these variables. The GAMS code for our case can then look as follows

```
*
* avoid branching on variables "chiS(t)" and "chiI(t)"
*
   chiS.prior(t) = +inf ;
   chiI.prior(t) = +inf ;

* use the branching priorities in the model
   portfolio.prioropt = 1 ;
```

Defining an arbitrary value > 1 for the branching priority for the variables χ_t^{S} and χ_t^{I} ensures that the branching on those variables is done only after all other variables have integral value. However, as the integrality of the variables δ_{mt} does not imply the variables χ_t^{S} and χ_t^{I} to be binary, it might be needed to branch on those variables nevertheless. One way where such a branching is not necessary is the case when there is a (non-zero) cost associated with the variables χ_t^{S} and χ_t^{I}; for instance, start-up cost for the power plan, see Sect. 4.2.

Carrion & Arroyo (2006) give a compact formulation of the minimum up and minimum idle time constraints using only one set of binary constraints – instead of two sets of variables χ_t^{S} and χ_t^{I}. However, they have a quadratic cost structure for the power plant and binary variables indicating if the power plant is used or not. Gröwe-Kuska et al. (2002) also use binary variables indicating if the plant is used in time slice t or not. Hence, they can also model the minimum up/down time requirement without using additional binary variables.

[4] Recognize that for this argument to be correct, we need also that the heuristics treat both the binary and the continuous case equivalently as well as factional solutions for the variables χ_t^{S} and χ_t^{I} are not rejected by the heuristics and during the branching process. However, just setting the branching priorities low, i.e. to value 10, has already a significant impact. For our case of the real data, the running time decreased by 30%.

4 Improvements of the Model Formulation

4.1 Assumptions and Limitations of the Model

Here, we discuss the assumptions needed for our model and present some limitations.

1. The pricing for the LFC is very simplified. In practice, there are special rebates; that is, they depend on the total energy purchased or the ratio of energy purchased to maximal power drained.
2. Although the electric power forecast is accurate enough for about a week, the increase of the time horizon to two or more days is computationally expensive and thus limits the application of this model.
3. As public services in Germany usually do not sell energy in the spot market, our model does not include this feature. Indeed, allowing to trade excess energy, leads to a different kind of optimization problem: One would operate the own power plant at an optimal efficient level and optimize the sale and purchase of the remaining/excess energy in the market.

   An overview of the behavior of such a market can be found in the book edited by Schweppe et al. (2002).

4.2 Modifications

- **EEG: Renewable Energy Act:** A law to regulate the priority of renewable energies in Germany; Bundesministerium für Wirtschaft und Technologie (2004, 2006). Especially the expansion of wind energy is intended. It forces electric distributors having wind-energy plants in their portfolio for their service area. Hence, it forces the additional purchase of wind-energy. However, the exact amount produced by wind is unpredictable. The optimization model has to treat this energy source stochastically. Stochastic optimization models and algorithms for this topic have been widely discussed in literature.
- **Hour Contracts:** The power bourse EEX also offers hour contracts which refer only to a specific hour. Those hour contracts can be used to fill up some small portion of the portfolio which is not covered by the base load and peak load contracts.
- **Emission Modeling:** The environmental issues in power generation play an important role. Especially the emissions of CO_2, NO_x or SO_x are currently under restriction. This can be modeled, for instance, via hard or soft constraints on the generated emissions or by minimizing the cost associated with those emissions. However, in the latter case, it is difficult to derive appropriate costs for the emissions. This problem is called *environmental dispatch problem*. More details can be found, for instance, in Talaq et al. (1994) and Yalcinoz & Köksoy (2007).

- **Efficiency Factor under Partial Load:** The efficiency factor of a power plant decreases when it is operated only under partial load. In particular, the variable costs are not constant through the whole power range. Hence, for each power stage, a separate cost has to be assumed. This is not so much a problem from the point of view of the mathematical modelling, but it is particularly difficult to get realistic data; that is, the cost coefficients. Let $C_{\mathrm{m}}^{\mathrm{PP}}$ be the variable cost in € per MWh for the power plant when operated in stage $m \in \mathcal{M}$, $m \geq 2$. If those data are available, then we can substitute the variable cost $C_{\mathrm{var}}^{\mathrm{PP}} \cdot e^{\mathrm{PP}}$ of the power plant in (7) by

$$\frac{1}{40} P_{\max}^{\mathrm{PP}} \sum_{t=1}^{N^{\mathrm{T}}} \sum_{m=2}^{N^{\mathrm{M}}} C_{\mathrm{m}}^{\mathrm{PP}} (m+2)\, \delta_{mt}.$$

  Recognize that we do not need any additional variables or constraints.
- **Start-up Cost for the Power Plant:** In (7), we stated that the cost of the power plant consists of fixed cost $C_{\mathrm{fix}}^{\mathrm{PP}}$ and variable cost $C_{\mathrm{var}}^{\mathrm{PP}}$ per MWh produced by the plant. Those fixed costs apply whether we use the power plant during this day or not. Such fixed costs can be, for instance, capital costs. However, it is more realistic, to have also start-up costs, which occur whenever the power plant is operated from an idle state. Those costs are typically fuel-costs for warming up. Let $C_{\mathrm{su}}^{\mathrm{PP}}$ be the start-up cost for the power plant. Then, we can add the following cost

$$C_{\mathrm{su}}^{\mathrm{PP}} \sum_{t=1}^{N^{\mathrm{T}}} \chi_t^{\mathrm{I}}$$

  to the cost of the power plant c^{PP} given in (7).
  Similarly, one could define shut-down cost for the plant. However, in this case, additional variables would be needed. Recognize that we can also include stage-switching cost, applying whenever the power plant changes its stage of operation.
- **Down-Time or Forced Operation of the Power Plant:** In practice, it could occur that the power plant has to be shut-down for some time period; for example, due to scheduled maintenance. This can be handled straight forward with our model by defining

$$\delta_{1t} = 1,$$

  for all time slices t where we want to force the plant to be in idle state. This condition implies for a given t that $\delta_{mt} = 0$ for all $m \in \mathcal{M}$, $m \geq 2$ according to constraint (25).
  This can be easily done in GAMS with the following code

```
*
* force the power plant to be shut down in time slice 't17'
* i.e. to be in idle state in time slice 't17'
*
 delta.fx('m1','t17') = 1 ;
```

The same idea can be used to force the power plant to operate in a certain stage $m \in \mathcal{M}$, $m \geq 2$ or just not to be in the idle stage. Recognize that in all cases, the number of binary variables in our model are reduced.

5 Computational Results

The optimization model is implemented in GAMS, version 22.7. The code is included in the GAMS (2009) model library with the model name `poutil.gms`. All computations are done with a Pentium Intel Centrino Dual 2.00 GHz with 1 GB RAM and Windows XP platform. In order to achieve computational results that are comparable, we use only one processor. We observed that with two processors, the speed-up time is almost linear in average.

A GAMS code to use multiple processors looks as follows

```
*
* for parallel use of cplex
*
* create file 'cplex.opt'
* and set the number of threads to 2
 $ onecho > cplex.opt
    threads 2
 $ offecho

* use the option file 'cplex.opt' for the 'energy' model
 energy.optfile = 1 ;
```

Using the real data for the Stadtwerke Saarlouis GmbH (2003), a German distributor, we get a (proven) optimal solution within 987 seconds. The computational details are given in the first row of Table 1 and the solution is plotted in Fig. 3. The total energy demanded is given in the area below the power demand forecast.

Table 1 shows computational results for different electric power demand forecasts. The basis are some real data for the power forecast. The new power forecast is

Table 1 Computational results for different demand forecasts. The first row are the real data and all other data are (uniform) randomly generated within an absolute difference of 2%

	Power Plant		Spot Market				LFC				
#	e^{PP}	c^{PP}	α	β	e^{SM}	c^{SM}	e^{LFC}	c^{LFC}	c	# Nodes	CPU
1	6,015.0	150,375.0	90	5	2,220	71,580	694.00	44,838	266,793.0	59,300	986.61
2	6,120.0	153,000.0	82	14	2,136	69,864	663.75	43,265	266,129.0	24,100	734.63
3	6,172.5	154,312.5	78	12	2,016	65,808	747.75	47,633	267,753.5	100,500	1678.67
4	6,045.0	151,125.0	90	0	2,160	69,120	728.25	46,619	266,864.0	50,000	992.30
5	6,142.5	153,562.5	82	10	2,088	67,896	726.00	46,502	267,960.5	53,800	1511.08
6	6,165.0	154,125.0	80	11	2,052	66,852	723.75	46,385	267,362.0	87,100	1225.34
7	5,977.5	149,437.5	94	0	2,256	72,192	713.75	45,865	267,494.5	53,300	1550.58
8	6,292.5	157,312.5	71	18	1,920	63,384	707.25	45,527	266,223.5	41,700	1020.03
9	6,202.5	155,062.5	79	11	2,028	66,084	714.75	45,917	267,063.5	48,400	2020.84
10	6,202.5	155,062.5	79	9	2,004	65,100	727.75	46,593	266,755.5	51,400	845.41

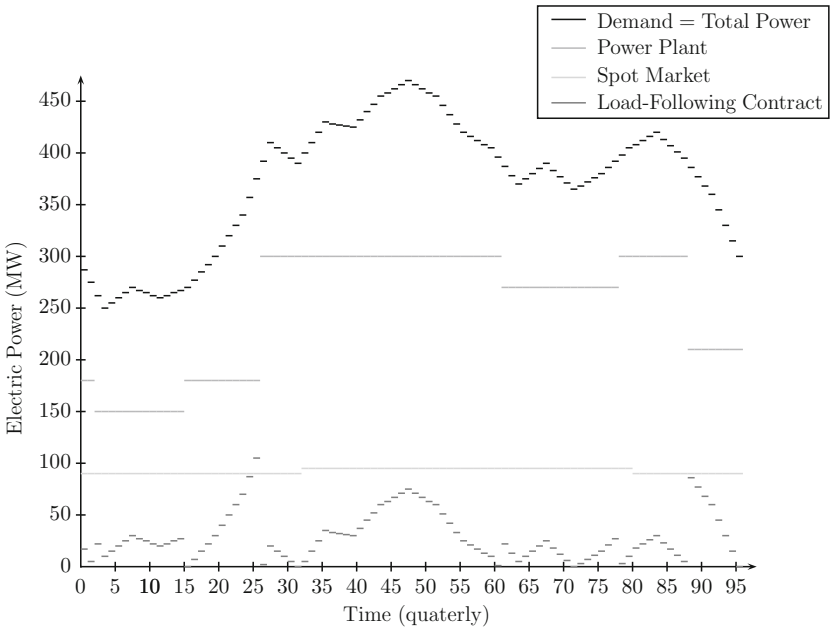


Fig. 3 Optimal solution for real data of Stadtwerke Saarlouis

randomly generated within a 2% tolerance. The column with label "# Nodes" gives the number of nodes in the Branch & Bound tree. The running time is stated in the last column and is measured in seconds. In all the 10 cases, the energy purchased from the LFC was enough to be in the cheapest price segment three. The borderline from price segment two to three is 500 MWh on a daily basis. Interestingly, the solutions differ quite remarkably when the energy forecast changes slightly; especially the purchase of energy from the spot market differ a lot.

In Table 2, the computational results for different minimum duration times between state changes of the power plant is shown. The power forecast are the same in all computations. We can observe that the change in the duration does not affect the solution very much. In fact, the difference in the total cost between a duration time of 1 h and 4 h is less than 2%. One explanation can be found in Fig. 3 as the power level of the power plant does not change every 2 h. Hence, a change in the duration has not such a big effect. As expected, the computational running time decreases when increasing the duration D^{PP}_{act}. An optimal solution for the duration of 4 h is shown in Fig. 4.

We performed also some computational tests for the case of a two-day planning horizon, $N^{\mathrm{T}} = 192$. The tested instance could not be solved to global optimality and after 10 h of computation time, the gap was still 5.99%.

Table 2 Computational results for different minimum duration times D_{act}^{PP} between state changes of the power plant

	Power Plant		Spot Market				LFC				
D_{act}^{PP}	e^{PP}	c^{PP}	α	β	e^{SM}	c^{SM}	e^{LFC}	c^{LFC}	c	# Nodes	CPU
4	6,112.5	152,812.5	90	5	2,220	71,580	596.50	39,768	264,160.5	1471,000	15039.56
6	6,075.0	151,875.0	90	5	2,220	71,580	634.00	41,718	265,173.0	165,700	1736.77
8	6,015.0	150,375.0	90	5	2,220	71,580	694.00	44,838	266,793.0	59,300	986.61
10	6,022.5	150,562.5	90	2	2,184	70,104	722.50	46,320	266,986.5	25,300	685.06
12	6,165.0	154,125.0	75	17	2,004	65,964	760.00	48,270	268,359.0	19,500	798.44
14	6,060.0	151,500.0	80	15	2,100	68,820	769.00	48,738	269,058.0	15,900	459.86
16	6,060.0	151,500.0	80	15	2,100	68,820	769.00	48,738	269,058.0	7,700	358.73

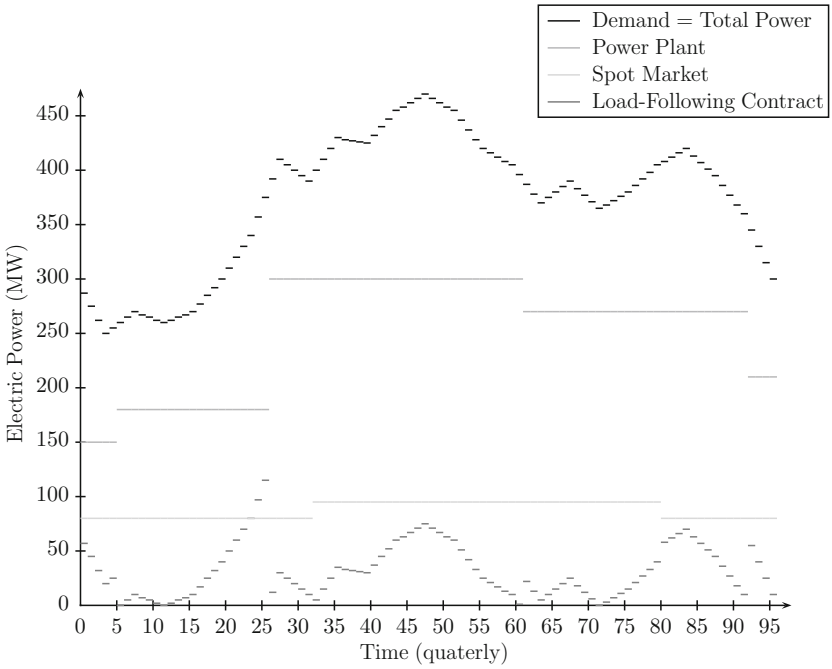


Fig. 4 Optimal solution for real data of Stadtwerke Saarlouis with $D_{act}^{\text{PP}} = 16$ (4 h)

6 Conclusion

In this article, we developed a model for the portfolio optimization of an electric services distributor. This study was motivated by a real case of public services in Germany. It brings together the real energy world and mathematical optimization. The model is very generic and can be easily extended with additional features but nevertheless, it has an appropriate degree of details matching the real world case. We also showed that the developed model is computationally effective for one-day

ahead planning. The developed model has also didactic value as some modelling tricks and their computational implications are discussed. The GAMS code is available in the GAMS (2009) model library.

Acknowledgement We would like to thank Peter Miebach (Mühlheim an der Ruhr, Germany) for providing the real-world case to us and his engagement in improving the description of the real world situation in this paper. Panos M. Pardalos and Steffen Rebennack are partially supported by AirForce and CRDF grants. The support is greatly appreciated.

References

Arroyo, J., & Conejo, A. (2000). Optimal response of a thermal unit to an electricity spot market. *IEEE Transactions on Power Systems, 15*(3), 1098–1104.

Atamtürk, A., & Savelsbergh, M. (2005). Integer-programming software systems. *Annals of Operations Research, 140*(1), 67–124.

Baldick, R. (1995). The generalized unit commitment problem. *IEEE Transactions on Power Systems, 10*(1), 465–475.

Beale, E., & Tomlin, J. (1969). Special facilities in a general mathematical programming system for non-convex problem using ordered sets of variables. In *5th International Conference on Operation Research* (pp. 447–454). North-Holland.

Brand, H., Weber, C., Meibom, P., Barth, R., & Swider, D. J. (2004). A stochastic energy market model for evaluating the integration of wind energy. In *Tagungsband der 6. IAEE European Conference 2004 on Modelling in Energy Economics and Policy*. Zurich.

Bruce, A. M., Meeraus, A., van der Eijk, P., Bussieck, M., Dirkse, S., & Steacy, P. (2009). McCarl GAMS User Guide. GAMS Development Corporation.

Bundesministerium für Wirtschaft und Technologie (2004). *Gesetz für den Vorrang Erneuerbarer Energien (Erneuerbare-Energien-Gesetz – EEG)*.

Bundesministerium für Wirtschaft und Technologie (2006). *Bundeskabinett beschließt Entlastungen im Erneuerbare-Energien-Gesetz (EEG)*. Berlin.

Carrion, M., & Arroyo, J. (2006). A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem. *IEEE Transactions on Power Systems, 21*(3), 1371–1378.

Chowdhury, S., & Rahman, B. H. (1990). A review of recent advances in economic dispatch. *IEEE Transactions on Power Systems, 5*(4), 1248–1259.

Dhillon, K., & Dhillon, J. S. (2004). "*Power system optimization*". India: Prentice Hall.

Dillon, T. S., Edwin, K. W., Kochs, H.-D., & Taud, R. J. (1978). Integer programming approach to the problem of optimal unit commitment with probabilistic reserve determination. *IEEE Transactions on Power Apparatus and Systems, PAS-97*(6), 2154–2166.

EEX - European Energy Exchange (2007). *EEX Product Information Power*. Leipzig.

Erdmann, G., & Zweifel, P. (2007). *Energieökonomik - Theorie und Anwendungen*. Berlin: Springer.

European Commission (2007). Germany - Energy mix fact sheet.

GAMS (2009). The GAMS model library index.

Gröwe-Kuska, N., & Römisch, W. (2005). Applications of stochastic programming. In S. W. Wallace & W. T. Ziemba (Eds.), *Stochastic unit commitment in hydro-thermal power production planning*, Chap. 30, MPS-SIAM Series in Optimization.

Gröwe-Kuska, N., Kiwiel, K. C., Nowak, M. P., Römisch, W., & Wegner, I. (2002). Decision making under uncertainty: energy and power. In C. Greengard, & A. Ruszczynski (Eds.), *IMA Volumes in Mathematics and its Applications* (Vol. 128, pp. 39–70), Power management in a hydro-thermal system under uncertainty by Lagrangian relaxation. New York: Springer.

Heuck, K., & Dettmann, K. -D. (2005). *Elektrische Energieversorgung: Erzeugung, Übertragung und Verteilung elektrischer Energie für Studium und Praxis* (6th ed.). Vieweg.

Hobbs, B., Stewart, W., Bixby, R., Rothkopf, M., ONeill, R., and Chao, H. -p. (2002). *The next generation of electric power unit commitment models, chapter why this book? New capabilities and new needs for unit commitment modeling* (pp. 1–14).

Kallrath, J., & Wilson, J. M. (1997). *Business optimisation using mathematical programming*. Houndmills, Basingstoke, UK: Macmillan.

LINDO Systems (2003). Application survey paper: electrical generation unit commitment planning.

Madlener, R., & Kaufmann, M. (2002). Power exchange spot market trading in Europe: theoretical considerations and empirical evidence. OSCOGEN Deliverable 5.1bMarch; Contract No. ENK5-CT-2000-00094.

Madrigal, M., & Quintana, V. (2000). An analytical solution to the economic dispatch problem. *IEEE Power Engineering Review, 20*(9), 52–55.

Nowak, M. P., & Römisch, W. (2000). Stochastic lagrangian relaxation applied to power scheduling in a hydro-thermal system under uncertainty. *Annals of Operations Research, 100*(1), 251–272.

Österreichische Elektrizitätsstatistikverordnung (2007). *284. Verordnung des Bundesministers für Wirtschaft und Arbeit über statistische Erhebungen für den Bereich der Elektrizitätswirtschaft*.

Padhy, N. (2004). Unit commitment-a bibliographical survey. *IEEE Transactions on Power Systems, 19*(2), 1196–1205.

Philpott, A., & Schultz, R. (2006). Unit commitment in electricity pool markets. *Mathematical Programming, 108*(2), 313–337.

Rosenthal, R. E. (1997). A GAMS Tutorial.

Rosenthal, R. E. (2008). *GAMS – A user's guide*. Washington, DC, USA: GAMS Development Corporation.

Schweppe, F. C., Caramanis, M. C., Tabors, R. D., & Bohn, R. E. (Eds.) (2002). *Spot pricing of electricty* (5th ed.). Boston, MA: Kluwer.

Sen, S., & Kothari, D. P. (1998). Optimal thermal generating unit commitment: a review. *International Journal of Electrical Power & Energy Systems, 20*(7), 443–451.

Sheble, G. B., & Fahd, G. N. (1994). Unit commitment literature synopsis. *IEEE Transactions on Power Systems, 9*(1), 128–135.

Shiina, T., & Birge, J. R. (2004). Stochastic unit commitment problem. *International Transactions in Operational Research, 11*(1), 19–32.

Stadtwerke Saarlouis GmbH (2003). *Jahreshöchstlast als viertelstündige Leistungsmessung*.

Takriti, S., Birge, J., & Long, E. (1996). A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems, 11*(3), 1497–1508.

Takriti, S., Krasenbrink, B., & Wu, L. S. -Y. (2000). Incorporating fuel constraints and electricity spot prices into the stochastic unit commitment problem. *Operations Research, 48*(2), 268–280.

Talaq, J. H., EI-Hawary, F., & EI-Hawary, M. E. (1994). A summary of environmental/economic dispatch algorithms. *IEEE Transactions on Power Systems, 9*(3), 1508–1516.

Verband der Netzbetreiber - VDN - e.V. beim VDEW (2006). *VDN-Richtlinie – MeteringCode 2006*. Berlin.

Wallace, S. W., & Fleten, S. -E. (2003). Stochastic programming. In A. Ruszczynski & A. Shapiro (Eds.), *Handbooks in operations research and management science, chapter Stochastic programming models in energy* (Vol. 10, pp. 637–677). North-Holland.

Wolsey, L. A., & Nemhauser, G. L. (1999). *Integer and combinatorial optimization*. Wiley-Interscience.

Wood, A. J. & Wollenberg, B. F. (1996). *Power generation, operation, and control* (2nd ed.). New York: Wiley.

Yalcinoz, T., & Köksoy, O. (2007). A multiobjective optimization method to environmental economic dispatch. *International Journal of Electrical Power & Energy Systems, 29*(1), 42–50.

A Indices and Index Sets

The indices, index sets and the indicator function of the mathematical programming model of Sect. 3 are given in the first column of Table 3. The second column states the name of the corresponding set/function used in the GAMS model `poutil.gms` included in the GAMS (2009) model library. The third column gives some explanations along with the size of the sets.

The model is generic and can tolerate in principle any number of time slices `Nt`. However, when changing the planning horizon, the modelling of the spot market has to be adjusted; for example, there has to be a variable α and β for each day of the planning horizon. In addition, the zones for the LFC have to be adjusted for the new horizon; for example, the step function I_t^{PL} in formula (10) has to be redefined.

In principle, the model can handle any number of power plant stages `Nm`. However, when changing this number, the formula for the power level p_t^{PP}, stated in (26), has to be changed too.

B Variables

All variables used in the mathematical model are summarized in the first column of Table 4. The corresponding variable name of the GAMS model `poutil.gms`, included in the GAMS (2009) model library, is given in the second column. A "-" in the second column states that this variable is not used in the GAMS model formulation, for example, the variable could be substituted by other variables. The units are stated in []-brackets in the third column and the forth column gives the type of the variable in the GAMS model formulation. $\mathbb{R}_+$, $\mathbb{Z}_+$, $\{0,1\}$ means that the variable is non-negative continuous, non-negative integer or binary, respectively. Recognize that this does not represent the domain of the variable but the type of the variable in the GAMS model. Particularly, the binary variables χ_t^{S} and χ_t^{I} are modelled being non-negative continuous; see Sect. 3.2.

Table 3 Indices, index sets and indicator function

$t \in \mathcal{T} := \{1,\ldots,N^{\mathrm{T}}\}$	`t`	Set of time slices per day. The day is split in N^{T} time intervals of 15 min each. $N^{\mathrm{T}} = 96$
$m \in \mathcal{M} := \{1,\ldots,N^{\mathrm{M}}\}$	`m`	Set for the power level stages of the power plant. The first stage corresponds to the stationary or idle phase of the plant; all other stages correspond to the 60–100% plant utilization stages. $N^{\mathrm{M}} = 8$
$b \in \mathcal{B} := \{1,\ldots,N^{\mathrm{B}}\}$	`b`	Set of support points of the zone prices for the LFC. $N^{\mathrm{B}} = 3$
I_t^{PL}	`IPL(t)`	Indicator function for the peak load contract. It is defined in (10)

Table 4 Variables with corresponding GAMS name, unit, model domain, equation reference(s) and explanations

Objective Function					
c^{tot}	`c`	[€]	$\mathbb{R}_+$	(6)	Total cost
Power Plant					
c^{PP}	`cPP`	[€]	$\mathbb{R}_+$	(7)	Cost associated with the power plant usage
e^{PP}	-	[MWh]	$\mathbb{R}_+$	(8)	Total amount of energy withdrawn from the power plant
p_t^{PP}	`pPP(t)`	[MW]	$\mathbb{R}_+$	(26)	Amount of power withdrawn from the power plant for time slice t. This variables can only have the discrete values 0, 0.6, 0.7, 0.8, 0.9 and 1.0 referred to the power plant capacity $P_{\max}^{\text{PP}}$
δ_{mt}	`delta(m,t)`	[−]	$\{0,1\}$	(24)	Binary variable with value 1 if the power plant is in time interval t in stage m and 0 otherwise
χ_t^{S}	`chiS(t)`	[−]	$\mathbb{R}_+$	(27), (28), (29)	Binary variable with value 1 if the power plant changes its stage at the beginning of time interval t and 0 otherwise
χ_t^{I}	`chiI(t)`	[−]	$\mathbb{R}_+$	(30)	Binary variable with value 1 if the power plant has been started up at the beginning of time interval t and 0 otherwise; that is, the power plant left the idle condition
Spot Market					
c^{SM}	`cSM`	[€]	$\mathbb{R}_+$	(12)	Cost for the energy purchase from the spot market
e^{SM}	-	[MWh]	$\mathbb{R}_+$	(11)	Energy purchased from the spot market
p_t^{SM}	`pSM(t)`	[MW]	$\mathbb{R}_+$	(9)	Electric power from the spot market for time slice t resulting from base load and peak load contracts
α	`alpha`	[−]	$\mathbb{Z}_+$		Quantity/proportion of the base load contracts of the portfolio contribution bought from the spot market. Typical range is between 0 and 200. We set as an upper bound the maximal demand in the planning horizon
β	`beta`	[−]	$\mathbb{Z}_+$		Quantity/proportion of the peak load contracts of the portfolio contribution bought from the spot market. Typical range is between 0 and 200. We set as an upper bound the maximal demand in the planning horizon
Load-Following Contract					
c^{LFC}	`cLFC`	[€]	$\mathbb{R}_+$	(14)	Cost for the energy purchase from load-following contract: energy rate
e^{LFC}	`eLFCtot`	[MWh]	$\mathbb{R}_+$	(16)	Total energy from the LFC
e_b^{LFC}	`eLFCs(b)`	[MWh]	$\mathbb{R}_+$	(20)	Contribution to the total energy of the LFC in segment b
p_t^{LFC}	`pLFC(t)`	[MW]	$\mathbb{R}_+$	(13)	Power from the LFC for time slice t
μ_b	`mu(b)`	[−]	$\{0,1\}$	(17)	Binary variables with value 1 if the daily purchased amount of energy lies between Z_{b-1}^{d} and Z_b^{d}

C Constraints

All constraints of the GAMS model `poutil.gms`, included in the GAMS (2009) model library, are summarized in the Table 5. The first column of Table 5 states the name of the constraint in the GAMS model, the second column gives the corresponding equation number of the mathematical programming formulation introduced in Sect. 3 and the third column gives a brief explanation.

D Input Data and Parameters

All input data/parameters of the mathematical model are stated in the first column of Table 6. The corresponding name of the GAMS model `poutil.gms`, included in

Table 5 Constraints of the GAMS model with corresponding equation number and explanations

Objective Function		
`obj`	(6)	Total cost
Power Demand		
`demand(t)`	(23)	Power demand constraint for each time slice t (quarter-hour)
Power Plant		
`PPcost`	(7)	Power plant cost. The fixed costs of the power plant are not included in this model
`PPpower(t)`	(26)	Power of power plant at time slice t
`PPstage(t)`	(25)	The power plant is in exactly one stage at any time slice t
`PPchiS1(t,m)`	(27)	Constraint on variable `chiS(t)` to track a stage change m at time slice t
`PPchiS2(t,m)`	(28)	Constraint on variable `chiS(t)` to track a stage change m at time slice t
`PPstageChange(t)`	(29)	At most, one stage change takes place within any $D_{\text{act}}^{\text{PP}}$ time slices
`PPstarted(t)`	(30)	Constraint on variable `chiI(t)` to indicate if the plant left the idle state at the beginning of time slice t
`PPidleTime(t)`	(31)	The idle time of the power plant has to last for at least $D_{\text{idl}}^{\text{PP}}$ time slices
Spot Market		
`SMcost`	(12)	Cost for the power from the spot market
`SMpower`	(9)	Power from the spot market
Load-Following Contract		
`LFCcost`	(21)	Cost for the power from the LFC as the energy rate
`LFCenergy`	(16)	Energy from the LFC for one day via LFC power
`LFCmu`	(18)	Constraint on the price segment
`LFCenergyS`	(19)	Energy from the LFC for one day via energy from the different segments
`LFCemuo`	(20)	Accumulated energy amount for the first segment
`LFCemug(b)`	(20)	Accumulated energy amount for all segments except the first one

Table 6 Input data/parameters with corresponding GAMS name, unit and explanations

Power Demand			
P_t	`PowerForecast(t)`	[MW]	Power demand forecast on a quarter-hour base
Power Plant			
$C^{\text{PP}}_{\text{fix}}$	–	[€]	Fix costs of the power plant
$C^{\text{PP}}_{\text{var}}$	`cPPvar`	[€/MWh]	Variable costs of the power plant
$P^{\text{PP}}_{\text{max}}$	`pPPMax`	[MW]	Power plant capacity in Megawatt
D^{PP}_{act}	–	[–]	Minimum number of time intervals between two consecutive stage changes of the plant. $D^{\text{PP}}_{act} = 8$. This is modelled in the GAMS code via the set `iS`
D^{PP}_{idl}	–	[–]	Minimum number of time intervals for the plant to remain in an idle period. $D^{\text{PP}}_{idl} = 16$. This is modelled in the GAMS code via the set `iI`
Spot Market			
C^{BL}	`cBL`	[€/MWh]	Cost per base load contract purchased
C^{PL}	`cPL`	[€/MWh]	Cost per peak load contract purchased
Load-Following Contract			
$C^{\text{LFC}}_{\text{PR}}$	–	[€/MW]	Cost for power rate; given in formula (5)
$C^{\text{LFC}}_{\text{PR,year}}$	–	[€/MWh]	Cost for power rate on an annual basis
$P^{\text{LFC}}_{\text{ref}}$	`pLFCref`	[MWh]	Electric power reference level for LFC
Z_b	`eLFCbY(b)`	[MWh]	Annual borders of quantity zones for LFC
Z^{d}_b	`eLFCb(b)`	[MWh]	Daily borders of quantity zones for LFC; $b \in \mathcal{B}$ and $Z^{\text{d}}_0 = 0$; calculated via formula (15)
P^{LFC}_b	`cLFCvar(b)`	[€/MWh]	Variable cost/price of LFC in segment b
C^{LFC}_b	`cLFCs(b)`	[€]	Accumulated variable cost of LFC up to segment b; calculated through (22)

the GAMS (2009) model library, is given in the second column; a "-" states that this variable is not used in the GAMS model formulation. The particular units of the data are given the []-brackets in the third column. Column four states some explanations as well as the value of the parameters. One instance, defining the values of the data, is given in the GAMS model `poutil.gms`.

The fixed costs $C^{\text{PP}}_{\text{fix}}$ of the power plant are not included in the model as they are irrelevant for the optimization decisions.

Investment in Combined Heat and Power: CHP

Göran Bergendahl

Abstract This study investigates the advantages of investing in plants for cogeneration, i.e., Combined Heat and Power (CHP), in case the heat is utilized for district heating. A focus is set on Swedish municipalities. The demand for heat is visualized in terms of load curves and duration diagrams. A standard diagram is chosen in order to analyze the dimensioning of a CHP plant. Two main alternative dimensions are analyzed in depth, namely to operate a plant with full capacity during eight months or alternatively during six months of the year. For each alternative, a CHP plant is compared to a heat water plant (a "boiler") and a biological fuel is compared to the one of natural gas. Then, further expansions are analyzed in a parametric way. The outcome is that it is efficient to choose the dimension so large that it will only be operating at full scale during three months of the year. It is also shown that CHP plant based on biological fuel is profitable and outstanding. These theoretical findings are then illustrated by data taken from 10 large Swedish municipalities – Göteborg, Helsingborg, Linköping, Lund, Malmö, Norrköping, Stockholm, Uppsala, Västerås, and Örebro. However, even if cogeneration is an energy efficient way to supply electricity and heat in these municipalities, there are constraints to invest. Examples are contracted deliveries of heat from outside, existing old plants, average cost pricing, and uncertainties in terms of future taxation principles.

1 The Economics of Cogeneration

Cogeneration or Combined Heat and Power (CHP) is a modern form of technology to produce simultaneously heat and power in the same plant and from the same energy source, such as oil, coal, natural gas, or biomass. In doing so, a substantial amount of energy may be saved in comparison to systems producing heat and power

G. Bergendahl
School of Business, Economics, and Law, University of Gothenburg, SE 405 30 Gothenburg, Sweden
e-mail: goran.bergendahl@handels.gu.se

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_15,
© Springer-Verlag Berlin Heidelberg 2010

separately. However, such a form of *joint production* requires an almost immediate use of the heat either in terms of district heating or as heat for industrial processes. Here, it is important to observe that heat for district heating rarely is transported over large distances, while the opposite is true for electricity. The effect has been that electricity is often sold over large networks and under competition, while district heating mostly is sold via local networks and under monopoly.

A cogeneration plant is viewed as being very energy efficient and friendly for the environment as it makes use of the steam after that it has passed through a turbine. Such a plant may reduce CO_2 emissions, power sector investments, and the cost of distribution to the end consumer (see e.g., IEA 2008, p. 4, 7, 25). However, to preserve such energy efficiency there must exist a demand for hot water either in terms of district heating or in terms of industrial processes.

Many different forms of cogeneration are nowadays in operation, such as steam backpressure turbines, steam condensing turbines, combined cycle gas/steam turbines, gas turbines with heat recovery, and internal combustion engines. Table 1 demonstrates that in 25 countries of the European Union (EU) about 10% of electricity production (299.2 TWh) came from cogeneration in the year 2002. Germany was the largest producer with 56.2 TWh. Denmark stood for the largest share (49.1%) of cogeneration out of the total generation of electricity while France (with 4.9%) was among the countries with the lowest shares. The corresponding data for Sweden in 2002 were 10 TWh or 6.8%. However, in Sweden substantial amounts of electric energy have been used for production of district heating from heat pumps making the net supply of electricity from cogeneration substantially smaller.[1]

Table 1 CHP generation and capacity for heat and electricity by specific countries 2002 (Danko & Lösönen 2006, Table 2)

Country[2]	EU-25	D	DK	F	S	SF
CHP Electricity Production (TWh)	299.6	56.2	19.3	28.6	10.0	22.6
Share of Total Electricity Generated (%)	9.1	9.8	49.1	4.9	6.8	38.0
CHP Electricity Capacity (GW)	91.6	26.4	5.4	6.5	3.2	5.8
CHP Heat Production (TJ'000)	2, 844.2	544.7	122.7	263.4	116.9	263.4
CHP Heat Capacity (GW)	236.1	48.7	10.6	23.0	7.5	15.4

[1] "Sweden has been successful in developing district heating. However, inside this sector only 4.7 TWh electricity is produced by cogeneration. At the same time, 4.2 TWh electricity is used for heating purposes. Consequently, only 0.5 TWh electricity remains to meet the demand for electricity outside the district heating sector." (Kommunförbundet et al. 2002, p. 6).

[2] D = Germany, DK = Denmark, F = France, S = Sweden, and SF = Finland.

Sweden has taken a unique position, as no other country inside EU has such a small percentage of the heat for district heating coming from cogeneration.[3] A question to pose is if this minor share for Sweden depends on lower profitability, on larger risks, or may be on higher taxes? Consequently, there is a large interest in finding out the profitability of cogeneration in general and for Sweden in specific. Therefore, the purposes of this paper are as follows:

1. To develop a procedure to evaluate investments in cogeneration in general, and to apply such a method on real world cases.
2. To compare the profitability and the environmental consequences of investing in cogeneration with an investment in a boiler plant for heat production only.
3. To investigate the potential for a set of larger Swedish municipalities to invest in cogeneration in order to serve their inhabitants with district heating.

This paper is based upon an earlier document written in Swedish (Bergendahl 2008). It is organized as follows. First, we analyze the demand for district heating formalized in terms of duration diagrams. Second, we present alternatives of investments in heat and power. Third, we estimate the profitability of alternative investment strategies. Fourth, we apply the best strategy on the largest municipalities in Sweden.

2 The Demand for District Heating and Its Duration

In Sweden, the annual demand for hot water for heating purposes has been estimated to about 70 TWh for residential buildings and to about 24 TWh for other buildings (STEM 2006). Out of these 94 TWh, district heating stands for approximately 44 TWh. About 300 municipalities in Sweden have systems for district heating. Approximately 50 of them make use of cogeneration for their production of hot water. The largest shares of district heating (85–90%) are found in big cities such as Göteborg, Linköping, Lund, Malmö, Norrköping, Stockholm, Uppsala, and Örebro.

The demand for district heating varies with the outdoor temperature. That implies that the demand is higher in the winter than in the summer and higher during the day than at night. The seasonal variations are often very large, which has been illustrated by Werner (1984). He has studied variations in the demand for heat day by day over the year 1978. He found a variation of approximately one to five between summer and winter. For a week in February 1979 he demonstrated an approximate variation of three to two.

The demand for electricity is more related to industrial production than district heating. Consequently, cogeneration may be used to produce more heat and less electricity during certain periods, and vice versa.

[3] Observe that in Denmark of 2003 CHP stood for a much higher share of the total electricity produced than in Sweden. A reason may be that in Sweden a substantial part of the electricity is produced by hydro power and nuclear power, while Denmark had to look for other electricity sources.

A *load diagram* is a basic instrument to analyze the demand for district heating (see Fig. 1). However, such a diagram cannot be directly used to determine the best combination of plants of different types. For that purpose the load diagram must be transformed into a *duration diagram* that is a diagram where the demand is ranked in the decreasing order over a year. A duration diagram may either express the amount of energy demanded (in GWh) during each time segment or the average power demanded (in MW) during the same time segments.

Duration diagrams (Fig. 2) are used to determine a portfolio of different plants to serve the demand for heat (and electricity). Plants with low running costs but with high fixed costs are useful for the base load. Plants with high running costs but low fixed costs are more useful for the top load.

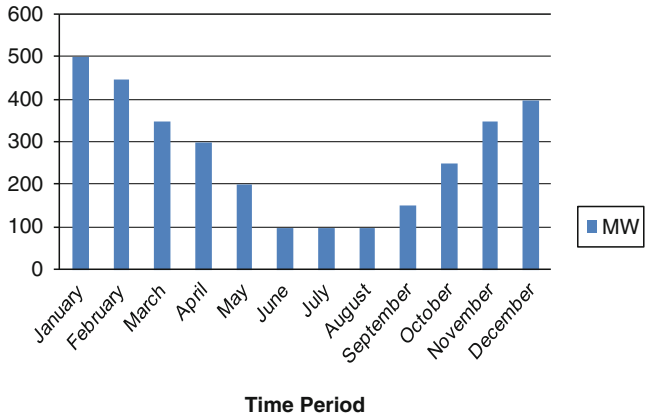


Fig. 1 Load diagram for the average power (MW) – example

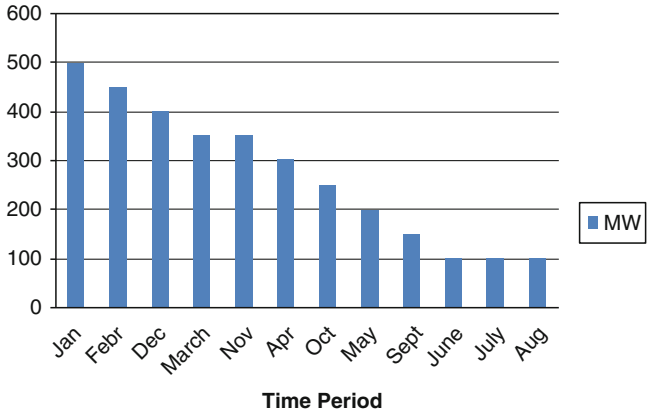


Fig. 2 Duration diagram for heat

Assume as given a duration diagram. Then a *normalized duration diagram* is obtained by dividing for each time segment the amount of energy demanded by the total annual demand. Normalized duration diagrams are useful in order to compare duration diagrams between different municipalities.

In general, duration diagrams are not officially available. However, a set of examples may be found in different sources such as the following ones from certain Swedish cities:

- Malmö 1979 (see Werner 1984)
- Piteå 1998 (see Byström 1999, p. 8)
- Uddevalla 2001 (see Johnsson & Rossing 2003, pp. 20–22)
- Varberg, Falkenberg and Halmstad 2003 (see Dahlberg-Larsson & Werner 2003, p. 15).

In these sources, the demand for heat has been given in terms of load curves. Below, these load curves have first been rearranged into duration curves, which in turn have been transformed into duration diagrams by splitting up the duration curves into 30-day time intervals (not necessarily equivalent to the monthly periods used in the example of Figs. 1, 2 above). The outcomes are shown in Table 2 below.

Then, these data have been normalized city by city by dividing them based on the corresponding annual demand for heat. These *normalized duration diagrams* are presented in Fig. 3 below.

In comparing these cities we observe that Malmö has a flatter duration diagram than the others, and that Piteå has the steepest slope on the duration diagram. That

Table 2 Average demand for heat over twelve 30-days periods of a year (in MW = MWh/h)

Period	1	2	3	4	5	6	7	8	9	10	11	12
Malmö 1979	489	438	390	333	333	291	243	195	150	100	100	100
Piteå 1998	3.6	2.8	2.5	2.2	1.9	1.7	1.4	1.1	0.9	0.7	0.5	0.4
Uddevalla 2001	47	44	42	37	34	28	25	17	14	9	7	7
Varberg, etc. 2003	205	150	140	130	120	110	90	70	60	45	35	25

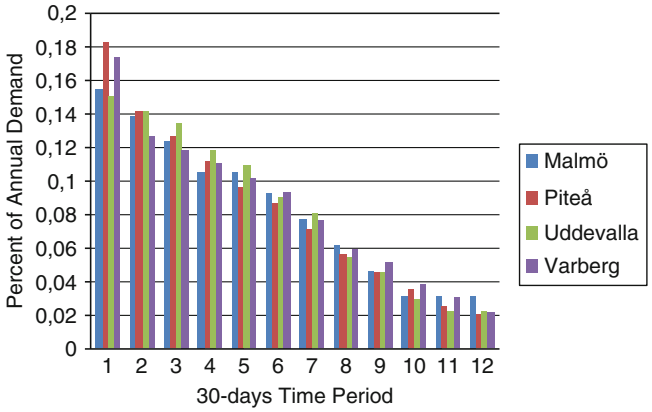


Fig. 3 Normalized duration diagrams for a set of Swedish municipalities

Table 3 A standardized distribution of the demand for heat, average capacity needed, and accumulated heat production over twelve 30-day periods of the year given a municipality with an annual demand of 1,000 GWh

30-Day period	Percent of annual demand	Average capacity needed (MW)	Accumulated production of heat (GWh)
1	16	244	178
2	14	213	334
3	13	198	478
4	11	167	600
5	10	152	711
6	9	137	811
7	7.5	114	894
8	6	91	961
9	5	76	1,016
10	3.5	53	1,055
11	2.5	38	1,083
12	2.5	38	1,110

Assume that the demand for heat per hour (MWh/h) is given for each of the 7,860 h of the year and presented in decreasing order that is as a duration diagram. First, take the 30 days with the highest demand and calculate the average capacity needed for that group. With 16% of an annual demand of 1,000 GWh, that implies $0.16 \times 1000/730 = 0.2192$ GWh/h or 219.2 MW. (Observe that $7,860/12 = 730$.) Finally, assume that we generate 10% losses in energy, why we have to produce $219.2/0.9 = 243.5$ MWh/h for the first 30-day period Then the same procedure is performed in consecutive order of the following eleven 30-day periods of the year resulting in the capacity needed as well as the accumulated production of heat Consequently, the third column states a duration diagram over the capacity needed for a municipality with a standardized demand. For example, the sixth 30-day period requires a production capacity of at least 137 MW and the eighth 30-day period a capacity of 91 MW.

indicates that the outdoor temperatures have a substantial influence on the shape of the duration diagrams. Malmö in the south have usually mild winters while Piteå in the north has cooler winters. Uddevalla, Varberg, Falkenberg, and Halmstad are all located along the Swedish west coast with about the same temperature conditions. The difference in the shape of the duration diagrams between these west coast cities may probably be explained by differences in temperature between different years. The data for Uddevalla were from the year 2001 while the data for Varberg, Falkenberg, and Halmstad were from the year 2003.

Below we will demonstrate a general procedure for calculating the profitability of investments in cogeneration compared to the one of a boiler. For the purpose of illustration we will introduce in Table 3 a "standardized" distribution of the demand for district heating in Sweden bearing in mind the fact that different municipalities will deviate from such a standard. Here, the standardization is based on the average demand for heat in the four studies of Swedish cities (see Table 2).

The example illustrated in Table 3 indicates a situation, where the first time period has an average production level of 244 MW and the second one an aver-

age level of 213 MW. Thus, it is relevant to say that the *duration* for 244 MW is 30 days; the duration for 213 MW is 60 days, etc. Consequently, a duration of one year is valid for 38 MW only. In principle, plants with high fixed costs but low running costs are suitable for a production with a long duration. On the other hand, the best use of plants with low fixed costs but high running costs is for shorter durations. Consequently, the flatter the duration diagrams the more use there will be for plants with high fixed costs but low running costs.

A cogeneration plant is a typical representative for plants with high fixed costs but low running costs. So, the flatter the duration diagram is the more use there is for cogeneration. Consequently, existing plants fuelled by oil or coal may become suitable for a production with a shorter duration, as their running costs are high but the fixed costs low (mainly because they have had a long period of depreciation).

Furthermore, an investment in cogeneration is usually associated with economies of scale, that is the larger the capacity the lower the average production cost. Therefore, large municipalities are expected to make more use of cogeneration than smaller ones may do.

An additional factor has to be taken into account. Many municipalities have signed long term contracts with industrial firms concerning deliveries of waste heat from their processes. In most cases, such contracts have duration of almost one year. Consequently, waste heat deliveries may stand for a substantial part of the "base load" capacity. That leads up to a production schedule such as demonstrated in Fig. 4. Therefore, a conclusion will be that the main keys to successful investments in cogeneration are rather *flat duration diagrams* with *few contracts for waste heat* from nearby industries.

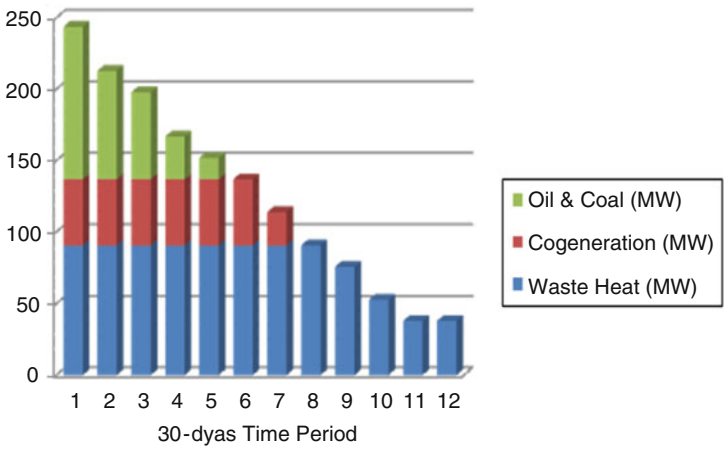


Fig. 4 An assumed production schedule for a fictitious municipality with an annual demand of 1,000 GWh district heating

3 Efficient Investments for the Production of Heat

There are two main ways to produce hot water for district heating. One is in a *boiler plant* (BP) and the other is a *cogeneration plant* (CHP). A cogeneration plant (Fig. 5) requires steam to be produced at a higher temperature than in a boiler plant. Consequently, *fuel costs* are higher for a cogeneration plant than for a boiler plant. Different fuels may be used, such as oil, coal, natural gas, forest products, and sewage. The choice of fuel will have a large influence on the running costs. Furthermore, investment costs are higher for a cogeneration plant than for a boiler plant, as the investments include a turbine to generate electricity.

In this section, we will analyze the best combination of investment and operation in order to deliver district heating in a municipality. With the "best combination" we first assume that environmental effects are met in terms of (a) the design of the investments, (b) operational constraints, and (c) environmental taxation. Given those conditions we will develop investment strategies being friendly for the environment and with the net present value as large as possible.

The environmental taxation is supposed to guide a municipality to choose fuels in order to reduce the emissions of substances that will hurt the environment. The estimation of these emissions given in Table 4 below is taken from STEM (2006).

In year 2007, Sweden introduced a new taxation system, with an aim to stimulate a reduction of these unwanted substances. That system is presented in Table 5 and the data are then inserted into the calculation schemes of Tables 6 and 7.[4]

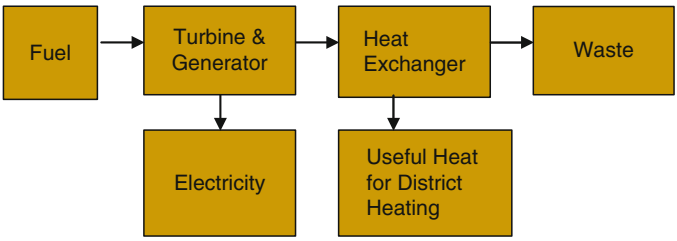


Fig. 5 Flow chart for a cogeneration plant

Table 4 Emissions of substances which may hurt the environment

Type of plant (Fuel)	Sulfur dioxid (g/MWh)	Nitrogen dioxid (g/MWh)	Carbon dioxid (kg/MWh)	Dust particles (g/MWh)	VOC[5] (g/MWh)
Oil	756	486	324	5.8	46.8
Coal	284	281	382	104.4	8.3
Waste	202	202	83	4.3	5.4
Biofuel	144	335	11	13.3	82.8
Natural Gas	12	237	209	1.2	12.6

[4] In addition to these taxes on the production and the use of energy, a customer has to pay the general value-added tax (VAT). However, the value-added tax will have no influence when comparing different investment alternatives.

[5] VOC = Volatile Organic Compounds.

Table 5 Swedish taxes on energy, carbon dioxide and sulfur as of year 2007

Fuel	Taxes on energy	Taxes on carbon dioxide	Taxes on sulfur dioxide	Total taxes
Oil, EO1	75	266	–	341
Oil, EO5	70	248	10	328
Coal (0.5% S)	43	315	20	378
Natural gas	20	185	–	205
Peat	–	–	15	15

Cogeneration is supposed to generate substantial advantages for the environment compared to power plants, for which the waste heat is disposed into the water without any productive use. Consequently, the Swedish parliament has decided to allow the owner of a cogeneration plant a *reduction on the taxes on energy and carbon dioxide* under the condition that the heat is given a productive use and not being wasted. That reduction will be 100% off the energy tax, 100% off the carbon dioxide tax when biomass will be the fuel, and 79% off the carbon dioxide tax when natural gas will be the fuel.

Furthermore, cogeneration has an advantage of *flexibility* compared to many other kinds of plants for power production. With cogeneration one may increase the proportion of heat and reduce the proportion of electricity when the electricity prices are low and reversely. Such an advantage may be handled as a *real option* and given a substantial value in the economic evaluation of an investment (see, e.g., Kulatilaka 2001; Cavus 2001; Olsson & Bergendahl 2006).

In the following, we will base our calculations on (a) the price on a certificate for power production is set to SEK 0.18/kWh, (b) the carbon dioxide tax has presently a level of SEK 0.378/kWh for natural gas from cogeneration and SEK 1.8/kWh for natural gas from a boiler plant, and (c) the energy tax of 0.2 SEK/kWh is only valid for the case of a boiler heated by natural gas.

Given these levels of pollution charges we will focus on investments that will take advantage of the economies of scale in cogeneration. Therefore, we will consider such a "standard" municipality, which expects an annual demand for heat of 1,000 GWh, mainly to be served by a new plant for operation in 20 years.

Furthermore, observe that we will assume that the use of cogeneration or alternatively a boiler is mainly to serve a municipality with the base load demand for heat. Consequently, we will aim at identifying an optimal load program for such a plant. That implies to find the number of months at full capacity leading to the largest net present value.[6] For purpose of illustration, we will start off with a situation for a municipality, which is going to choose between the following two alternative strategies for the design of a base load investment.

- *Strategy 1*: The maximum plant size is designed for full scale utilization in eight months of a year.

[6] It will be assumed that the investment costs for any plant will occur at the beginning of year 1 and that the annual costs and revenues from its operation will be charged at the end of each of the 20 years of operation.

Table 6 Four investment alternatives for a plant to be operated at full capacity during eight months of a year

Plant	CHP	CHP	Boiler	Boiler
Fuel	Natural Gas	Biomass	Natural Gas	Biomass
Capacity heat (MW)	91.3	91.3	91.3	91.3
Capacity electricity (MW)	102.6	33.6	–	–
Production heat (GWh/year)	533.2	533.2	533.2	533.2
Production electricity (GWh/year)	599.2	196.2	–	–
Fuel needed (GWh/year)	1,132.4	729.4	533.2	533.2
Investment (MSEK)	750	750	130	300
Maintenance (MSEK/year)	15	24	3	7.5
Fuel price (SEK/MWh)	225[7]	150[8]	225	150
Taxes CO_2 (SEK/MWh)[9]	37.8	–	180	–
Fuel costs (MSEK/year)[10]	297.6	109.4	215.9	80.0
Emission rights (MSEK/year)[11]	22.3	–	10.5	–
Electricity price (SEK/kWh)[12]	0.43	0.43	–	–
Electricity certificates (SEK/kWh)[13]	–	0.18	–	–
Revenues electricity (MSEK/year)	231.9	119.7	–	–
Heat price, excl. distr. (SEK/kWh)[14]	0.5	0.5	0.5	0.5
Energy taxes (SEK/kWh)	–	–	0.2	–
Revenues heat (MSEK/year)[15]	239.9	239.9	144.0	239.9
Net revenues (MSEK/year)[16]	136.9	226.2	−85.4	152.4
Net present value (MSEK)	956	2,068	−1, 194	1,599

[7]The price level for natural gas is estimated to SEK200–250/MWh.

[8]This price level seems relevant for splinters. The price level for pellets is assumed to be somewhat higher or about SEK170/MWh.

[9]In Sweden, a boiler heated with natural gas have to pay 100% carbon dioxide tax (SEK180/MWh), while a CHP plant heated by natural gas only have to pay 21% of the carbon dioxide tax (i.e. 0.21× SEK180/MWh).

[10](Fuel Needed) × (Fuel Price + Taxes).

[11]Assume an emission of 201 tons CO_2 per GWh fuel and a fuel price of SEK98/ton fuel (see Särnholm 2005, p. 37).

[12]See www.nordpool.com for statistics on energy prices 2006. They indicate that the eight highest monthly prices had an average of SEK0.43/kWh. See also Fastigheten Nils Holgersson 2006, Fig. 3, p. 14.

[13]Electricity certificates are usually not valid for more than 15 years. However, we assume here that they may be prolonged for five additional years.

[14]The price level for heat including the costs of distribution is usually set to SEK0.6–0.7/kWh (e.g. see Fastigheten Nils Holgersson 2006, pp. 24–25). The costs of distribution are analyzed by Frederiksen & Werner 1993, pp. 372–376. They assume that the total costs of distribution are about SEK0.04–0.05/kWh. In addition, they assume costs for losses in pressure. Consequently, SEK0.5/kWh seems to be a relevant level for the market price exclusive distribution.

[15](Heat Production) × 0.9 × (Heat Price -Taxes). The factor 0.9 stands for the amount of heat sold in relation to the amount produced.

[16](Revenues Electricity) + (Revenues Heat) − (Fuel Costs) − (Emission Rights) − (Maintenance).

Table 7 Four investment alternatives for a plant to be operated at full capacity during six months per year

Plant	CHP	CHP	Boiler	Boiler
Fuel	Natural Gas	Biomass	Natural Gas	Biomass
Capacity heat (MW)	137	137	137	137
Capacity electricity (MW)	153.1	49.8	–	–
Production heat (GWh/year)	749.9	749.9	749.9	749.9
Production electricity (GWh/year)	838.1	290.8	–	–
Fuel needed (GWh/year)	1,588.0	1,040.7	749.9	749.9
Investment (MSEK)	1,100	1,080	190	440
Maintenance (MSEK/year)	22	34	4.5	11
Fuel price (SEK/MWh)	225	150	225	150
Taxes CO_2 (SEK/MWh)	37.8	–	180	–
Fuel costs (MSEK/year)	417.3	156.1	303.7	112.5
Emission rights (MSEK/year)	31.3	–	20.5	–
Electricity price (SEK/kWh)	0.434	0.434	–	–
Electricity certificates (SEK/kWh)	–	0.18	–	–
Electricity revenues (MSEK/year)	327.0	160.7	–	–
Heat price excl. distr. (SEK/kWh)	0.5	0.5	0.5	0.5
Energy taxes (SEK/kWh)	–	–	0.2	–
Revenues heat (MSEK/year)	337.5	337.5	202.5	337.5
Net revenues (MSEK/year)	194.3	308.1	−126.2	214.0
Net present value (MSEK)	1,321	2,759	−1, 762	2,226

- *Strategy 2*: The maximum plant size is designed for full scale utilization only in six out of eight months of a year.

Then, the remaining part of the demand for 1,000 GWh is supposed to be met through capacities already in operation.

Four investment alternatives will be displayed for each of the two strategies above.

1. A cogeneration plant to be fuelled by biomass.
2. A cogeneration plant to be fuelled by natural gas.
3. A boiler to be fuelled by biomass.
4. A boiler to be fuelled by natural gas.

Finally, if Strategy 2 is found to be the most profitable one for all the four cases, we will go further on to investigate if cogeneration should be designed for a full scale utilization in a fewer number of months than the six ones.

3.1 The Economics of Strategy 1

In Table 6, we will analyze the benefits and costs associated with the above mentioned four alternatives based upon an assumption that the scale is chosen in order to operate at capacity in eight months. That leads up to the following assumptions.

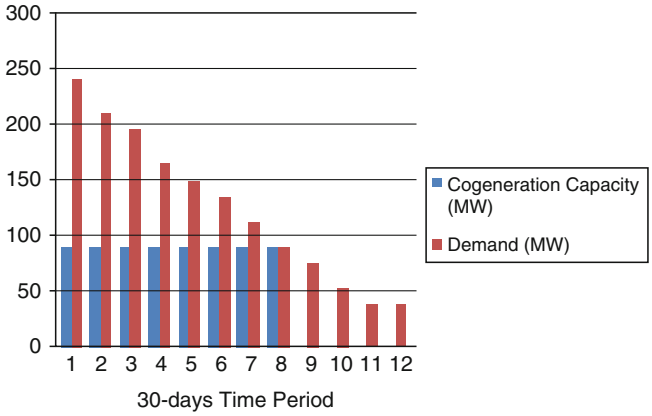


Fig. 6 A duration diagram for a cogeneration plant designed to meet the base load with full capacity during eight months

- All four alternatives are dimensioned in order to produce 91.3 MW heat (see Fig. 6 above).
- In case natural gas is to be used as the fuel, the cogeneration plant may get a power capacity of 102.6 MW. Alternatively, if biogas is treated as the alternative, the capacity for power will only become 33.6 MW.
- The investment cost for a cogeneration plant is estimated to SEK750 million for both the case with natural gas and the one with biomass. In case biomass is chosen as the fuel, the exhaust gases will be used to extract additional energy for district heating.
- The maintenance costs are estimated to SEK15 million/year for natural gas and SEK24 million/year in case of biomass.
- The fuel costs are estimated to SEK225/MWh for natural gas and SEK150/MWh for biomass.

Given these assumptions we may calculate the net present values for the four alternatives in accordance with Table 6 above. The discount rate will be set to 5% in general.

3.2 The Economics of Strategy 2

Strategy 2 is based on an assumption that a municipality with the same annual demand for district heating of 1,000 GWh/year may find it profitable to expand the cogeneration capacity to a level above what is needed for a full scale production that is above the 91.3 MW heat. In order to investigate the economics of such an expansion, we will focus on a capacity which allows for full scale utilization only in six out of eight months. For the remaining two months production is assumed

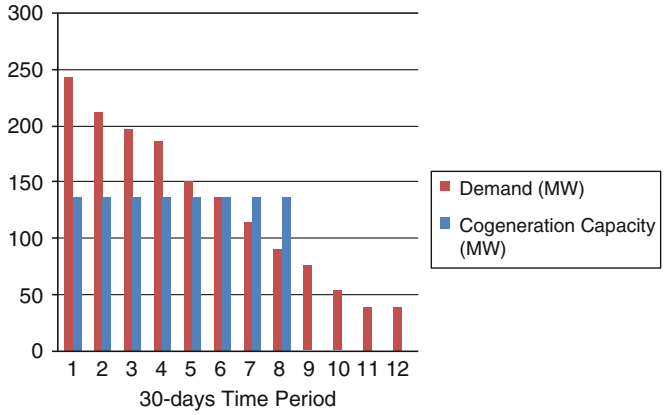


Fig. 7 A duration diagram for a cogeneration plant designed to meet the base load with full capacity during six out of eight months

at a level just to meet demand. Table 3 demonstrates that one may then expand the production capacity to 137 MW, which is illustrated by Fig. 7 above.

As a consequence, the investment costs for cogeneration will increase to a level of 1,100 MSEK for a plant based on natural gas and to about 1,080 MSEK for the case of biomass as a fuel. If one considers an investment in a boiler, then the investment costs are supposed to become substantially lower, that is 190 MSEK, when natural gas is supposed to become the fuel or, alternatively, 440 MSEK in case biomass is to become the fuel. Such a plant will be run at capacity (137 MW) in six months of the year. For the seventh month the production level will stay at about 114 MW and for the eighth month at about 91 MW (compare Table 3).[17] The outcome is given in Table 7.

3.3 Results from the Profitability Analyses

In Sects. 3.1 and 3.2, we have calculated the profitability of investing in a cogeneration plant as well as in a boiler plant. Both of them were designed for a base-load production during eight months per year. The peak load production during these eight months as well as the off-peak production during the remaining four months was supposed to be served by existing capacities.

Strategy 1 assumed an investment of 91.3 MW, which will allow for a full capacity use in eight months of the year. Thus, we may identify such an investment as one with *a base load factor of 0.0913,* that is a 91.3 MW investment per 10^6MWh

[17] It is assumed that the Swedish systems of energy taxation and regulation will prevent such an overcapacity in the seventh and eighth month to be used for an unilateral production of electricity (see the note above on a 100% tax reduction).

Table 8 Computational results

Alternative	Capacity heat (MW)	Capacity power (MW)	Production heat (GWh/year)	Production electricity (GWh/year)	Investment (MSEK)	Net present value (MSEK)
Strategy 1						
COGEN/NG	91.3	102.6	533.2	599.2	750	956
COGEN/Bio	91.3	33.6	533.2	196.2	750	2,068
Boiler/NG	91.3		533.2		130	−1,194
Boiler/Bio	91.3	–	533.2	–	300	1,599
Strategy 2						
COGEN/NG	137	153.1	749.9	838.1	1,100	1,321
COGEN/Bio	137	49.8	749.9	290.8	1,080	2,759
Boiler/NG	137	–	749.9	–	190	−1,762
Boiler/Bio	137	–	749.9	–	440	2,226

(1,000 GWh) annual heat demand. Strategy 2 went for an investment of 137 MW, which will only allow for a full capacity use in six out of eight months. However, it will result in a larger production volume than Strategy 1. As for fuel, we have considered two choices – natural gas or biomass. That ended up in eight different alternatives, for which the outcomes are presented in Table 8 above.

The calculations demonstrate the following findings.

- A cogeneration plant is more profitable than a boiler plant.
- Biomass is a better fuel alternative than natural gas (at least under the present system of taxation).
- It is profitable to design the capacity at least for a size of 137 MW (Strategy 2).

However, the question still to be answered is whether it is efficient to expand the capacity above 137 MW.

3.4 A Parametrical Expansion of the Capacity

In order to answer the question of an expansion above 137 MW, let us make a rough estimation by starting out from Strategy 2, alternative "*COGEN/Bio*" with a capacity of 137 MW heat. Then, make the following assumptions.

1. The additional cost for expanding the size of the investment above 137 MW will become (1,080–750)/(137–91.3) MSEK per MW = 7.22 MSEK per MW. That is, we assume that the marginal investment costs are proportional to the additional size.
2. The additional net revenues from such an expansion are then assumed to become proportional to the sales of heat. With the net revenues of 308.1 MSEK for a sales volume $0.9 \times 749.9 = 674.9$ GWh/year which implies SEK0.457/kWh.
3. The discount rate is set to 5%. For a horizon of 20 years that implies a discount factor of 12.46.

Then, follow Table 3 and consider a stepwise expansion from 137 to 152, 167, 198, 213, and 244 MW, corresponding to a full scale utilization in 5, 4, 3, 2, and 1 months, respectively. The outcomes in terms of net present values are presented in Table 9 below.

Table 9 demonstrates that given the assumed data in terms of costs and revenues it seems optimal to expand the cogeneration capacity above the 137 MW capacity of Strategy 2 to a level of 213 MW but not as far as 244 MW. The outcome of such a decision is illustrated in Fig. 8. It shows the strategy of investing in a cogeneration plant fuelled by biomass to be operated with full capacity in two out of eight months. However, such a conclusion is very sensitive to the level of the investment cost as well as the level of sales. For example, Table 9 shows that a 4% increase in the investment outlays a 4% reduction in the sales revenues will result in that the case of 2 months full scale capacity utilization only will not be profitable.

Then the corresponding sales volumes of heat from cogeneration will be expanded from 674.9 GWh/year to (674.9 + 49.3 + 39.4 + 61.1 + 19.7) = 844.4 GWh/year and the sales volumes of electricity from 261.7 GWh/year to (261.7 + 19.1 + 15.3 + 17.0) = 320.6 GWh/year. That gives us the following "Standard Expansion Module" to be considered by any large Swedish municipality.

A Standard Expansion Module
Annual Heat Demand: 1,000 GWh
Optimal Cogeneration Capacity: 213 MW
Fuel: Biomass
Cogeneration in Operation: 8 months/year
Optimal Sales from Cogeneration:
• Heat: 844.4 GWh/year
• Electricity: 320.6 GWh/year

Consequently, the Standard Expansion Module is designed for sales of 1,000 GWh heat out of which 84% should come from cogeneration. Below we will show how ten large Swedish municipalities may realize these potential benefits from cogeneration.

4 Potential Investments in Cogeneration for 10 Large Swedish Municipalities

In this section, we will investigate the potential for profitable investments in cogeneration in larger Swedish municipalities. To allow for economies of scale we will focus on the 10 municipalities with the largest demand for district heating.[18] They are Stockholm, Göteborg, Malmö, Uppsala, Västerås, Linköping, Örebro, Norrköping,

[18] It is evident that cogeneration may be profitable for smaller municipalities as well. However, this paper will not focus on finding a break-even point for cogeneration.

Table 9 Net Present Values of an additional expansion of the capacity above 137 MW[19]

(a) Full scale utilization (Months)	(b) Capacity (MW)	(c) Additional investment (MW)	(d) Additional investment outlays (MSEK)	(e) Additional sales of heat (GWh/year)	(f) Additional net revenues (MSEK/year)	(g) NPV of additional net revenues (MSEK)	(h) Total additional NPV (MSEK)
5	152	15	108.3	49.3	22.5	280.4	172.1
4	167	15	108.3	39.4	18.0	224.3	116.0
3	198	31	223.8	61.1	27.9	347.6	123.8
2	213	15	108.3	19.7	9.0	112.1	3.8
1	244	31	223.8	20.3	9.3	115.9	−107.9

[19] Column (b) gives the capacities corresponding to the duration diagram of Table 3. Column (c) presents the investments needed to expand the capacity. Column (d) stands for the corresponding investment costs. Column (e) gives the corresponding expansion in sales of heat based upon the data for the expansion from 91.3 to 137 MW. Column (f) states the expansion in net revenues based upon the data for the plant capacity 137 MW. Column (g) presents the net present values (NPV) of the net revenues over 20 years. Finally, Column (g) states the "Total Additional NPV" as the difference between Column (g) and Column (d).

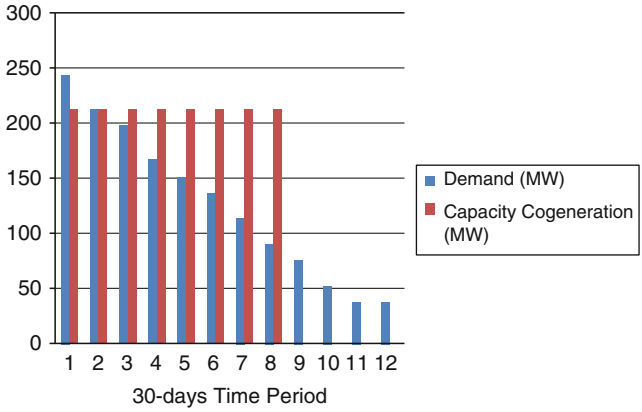


Fig. 8 A duration diagram for a cogeneration plant designed to operate with full capacity during two out of eight months

Table 10 Estimated potentials for cogeneration (COGEN) in the ten largest Swedish municipalities in year 2003[20]

(a) Municipality	(b) Deliveries of District Heating 2003 (GWh)	(c) Actual Heat from COGEN 2003 (GWh)	(d) Heat from COGEN by Use of the Strategy for a Standard Module (GWh)	(e) COGEN Capacity by Use of the Strategy for a Standard Module (MW)
Stockholm C	2, 916	1, 347	2,462#	621
Stockholm V	1, 038	966	876	221
Stockholm S	2, 968	1, 310	2,506#	632
Uppsala	1,596	844	1,348#	340
Linköping	1,280	1,386	1,081	272
Norrköping	948	1,024	800	202
Malmö	2,257	871	1,905#	481
Lund	875	207	739#	186
Helsingborg	913	632	771#	194
Göteborg	3,450	437	2,913#	735
Örebro	1,038	813	876#	821
Västerås	1,459	1,627	1,232	311

[20] In 2003, Stockholm operated three separate systems for district heating, Central (C), West (V), and South (S).

Helsingborg, and Lund. Based on available statistics from year 2003 it has been possible to estimate the potentials for investment in cogeneration in these cities.

Table 10 presents these potentials. The actual demand for district heating as of 2003 is given in column (b) and the corresponding cogenerated heat of that year in column (c). Then we estimate the amount of heat that would come from

cogeneration in case we use the above optimal strategy for a Standard Expansion Module (column (d)). Finally, the potential cogeneration capacity is given in column (e). For example, Linköping delivered 1,280 GWh heat during the year 2003. Then the Standard Expansion Module recommended a use of 213 MW cogeneration for each sales of 1,000 GWh. That is, we will arrive at a recommendation for Linköping to operate with $213\,\text{MW} \times 1,280/1,000 = 272.6\,\text{MW}$ cogeneration. Then the annual sales of heat from cogeneration will become GWh $844.4 \times 1,280/1,000 = 1,081$ GWh.

Let us now investigate Table 10. Compare the actual production of heat from cogeneration (column (c)) with the ones obtained by the use of a Standard Expansion Module (column (d)). Then we will find that in 2003 there seems to have been a potential for cogeneration investments in Stockholm C, Stockholm S, Uppsala, Malmö, Lund, Helsingborg, Göteborg, and Örebro (all marked with #). However, some of them have already signed long term contracts for purchasing of waste heat from industries and other organizations. For example, E-ON in Malmö will purchase 1,150 GWh/year from the two firms NCB and SYSAV, Öresundskraft in Helsingborg has signed up for 340 GWh/year in terms of industrial waste, and Göteborg Energi has contracted 1,000 GWh/year from the waste destruction firm Renova and 1,000 GWh/year in terms of industrial waste heat from Shell and Preem. Vattenfall Värme in Uppsala has invested substantially in a waste destruction plant, which will reduce the need for cogeneration substantially. That leads us to a conclusion that among the above 10 municipalities, the main potentials for investments in cogeneration seems to be found in Stockholm and in Lund.

In reality, since in 2003 the municipality of Stockholm investigates an investment in cogeneration by biomass, Lund looks for a biomass investment of 50 or 100 MW, Malmö goes for a new plant based on natural gas, and finally, Göteborg has been invested in a new plant based on natural gas. The actions taken by Göteborg and by Malmö may be explained by the fact that their potentials for cogeneration seem to be so large (2,913 GWh and 1,905 GWh, respectively, as given in Table 10) that they allow both for external deliveries of heat and for cogeneration production.

5 Limits for Investments in Cogeneration

In this paper, we have demonstrated a procedure to evaluate potential investments in cogeneration. In Sect. 3, it has been shown that an investment in cogeneration is more efficient than an investment in a boiler plant. Simultaneously, we have also found that it seems economically motivated to invest in a capacity so large that full capacity utilization is only expected during two out of eight months of operation. Furthermore, given a set of reasonable assumptions on prices and taxes in Sweden, it has been shown that biomass is a better fuel alternative than natural gas. However, the larger the investment costs per unit of capacity and the lower the price level for electricity and heat, the smaller the optimal size of investment. Then two months of full capacity utilization may not be enough.

Furthermore, there are limits for investment. The following ones seem to be most important ones.

- Contracted deliveries of waste heat from industries and sewage destruction plants reduce the potentials for cogeneration.
- Existing capacities in terms of old plants based upon the use of oil, coal, and other non-renewable fuels may postpone new investments in cogeneration.
- The slope of the load curve will influence the profitability of cogeneration. The flatter the load curve is, the more efficient the use of cogeneration will become.
- A use of average cost pricing will delay the investment in cogeneration. Lower prices in the summer time combined with higher prices in the winter time will instead stimulate a more even use of district heating and thus clear the way for more cogeneration.
- The uncertainty of the future taxation principles may deter firms from investing. Lower taxes on district heating compared to individual heating systems may be motivated both for environmental reasons and for economic purposes.
- Today, cogeneration is exempted from energy taxation except for any time period when it is used for electricity generation only or for heat generation only. As it may well be efficient to run periodically a CHP as a single product plant, those exceptions ought to be eliminated.

However, in spite of these limitations it seems reasonable to assume that cogeneration ought to become one of the most important means to meet a growing future demand for electricity and district heating. The given analysis shall be seen as a procedure to calculate how efficient cogeneration is compared to other alternatives of producing heat and electricity.

Acknowledgements Thanks go to the following persons, who have given me assistance in writing this report:

– Margaret Armstrong, Ecole des Mines, Paris
– Mats Bärring, E.On Värme Sverige AB, Malmö
– Bengt-Göran Dalman, Göteborg Energi AB, Göteborg
– Erik Dotzauer, AB Fortum Värme, Stockholm
– Anders Eriksson, Mälarenergi AB, Västerås
– Alain Galli, Ecole des Mines, Paris
– Ingvar Karlsson, Tekniska Verken AB, Linköping
– Tom Kerr, International Energy Agency, Paris
– David Knutsson, Göteborg Energi AB, Göteborg
– Christer Olsson, Öresundkraft Produktion AB, Helsingborg
– Stefan Persson, Vattenfall AB Värme, Uppsala
– Nils-Ove Rasmusson, Eslöv Lund Kraftvärmeverk AB, Lund
– Sigrid Sjöstrand, University of Lund, Lund
– Ulrik Snygg, Öresundkraft Produktion AB, Helsingborg
– Ulrik Stridbaek, International Energy Agency, Paris
– Peter Svahn, University of Gothenburg, Göteborg
– Sven Werner, Halmstad University, Halmstad
– Ove Åsman, AB Fortum Värme, Stockholm

as well as to two anonymous referees.

References

Bergendahl, G. (2008). *Investeringar i kraftvärme – Ekonomiska och miljömässiga fördelar ("Investment in Cogeneration – Advantages for the Economy and the Environment"). FE Report 2008–413.* , Göteborg: School of Business, Economics, and Law, University of Gothenburg.

Byström, R. (1999). *Närvärme i Norrfjärden ("Local Heating in the Norrfjärden"). Report 1999:40.* Lund Technical University.

Cavus, M. (2001). Valuing a power plant under uncertainty. In S. Howell et al. (Eds.), *Real options. evaluating corporate investments in a dynamic world*. Pearson Education.

Dahlberg-Larsson, C., & Werner, S. (2003). *Förstudie Regional fjärrvärme Halland ("Prestudy of District Heating in Halland").* Borås: FVB, Analysgruppen Borås, 2003–02–28.

Danko, J., & Lösönen, P. (2006). Statistics in focus, environment and energy. *Eurostat*, *3*, 1–8.

Fastigheten Nils Holgerssons underbara resa genom Sverige (2006). *En avgiftsstudie för år 2006 ("The property Nils Holgersson's wonderful journey in Sweden").* Stockholm: Fastighetsägarna, Hyresgästföreningen, Riksbyggen, SABO & HSB.

Frederiksen, S., & Werner, S. (1993). *Fjärrvärme. Teori, teknik och funktion, ("District heating, techniques, functioning").* Lund: Studentlitteratur.

IEA - International Energy Agency (2008). *Combined heat and power. Evaluating the benefits of greater global investment*. Paris.

Johnsson, J., & Rossing, O. (2003). *Samverkande produktions- och distributionsmodeller ("Cooperative models for production and distribution"), FOU 2003:83*. Stockholm: Fjärrvärmeföreningen.

Kommunförbundet, S., Fjärrvärmeföreningen, S., & Energi, S. (2002). *Tid för kraftvärme, ("Time for Cogeneration").* Stockholm.

Kulatilaka, N. (2001). The value of flexibility: the case of a dual-fuel industrial steam boiler. In Schwartz & Trigeorgis (Eds.), *Real options and investment under uncertainty*. MIT.

Olsson, K. O., & Bergendahl, G. (2006). *Investment in cogeneration – new methods to evaluate flexibility*. School of Business, Economics, and Law, Göteborg University.

Särnholm, E. (2005). *Åtgärdskostnader för minskning av koldioxidutsläpp vid kraftvärme- och värmeanläggningar, ("Costs for Reducing the Emissions of Carbon Dioxides at Cogeneration and District Heating").* Svenska Miljöinstitutet.

STEM (2006). *Värme i Sverige år 2005, ("Heating in Sweden in 2005").* Eskilstuna: Statens energimyndighet (The Government Energy Agency).

Werner, S. (1984). *The heat load in district heating systems*. Göteborg: Department of Energy Technology, Chalmers School of Technology.

Capacity Charges: A Price Adjustment Process for Managing Congestion in Electricity Transmission Networks

Mette Bjørndal, Kurt Jörnsten, and Linda Rud

Abstract In this paper, we suggest a procedure based on capacity charges for managing transmission constraints in electricity networks. The system operator states nodal capacity charges for transmission prior to market clearing. Market clearing brings forth a single market price for electricity. For optimal capacity charges the market equilibrium coincides with that of optimal nodal pricing. Capacity charges are based on technical distribution factors and estimates of the shadow prices of network constraints. Estimates can be based on market information from similar congestion situations, and then capacity charges can be brought near the optimal values through an iterative process.

1 Introduction

The goal of deregulating the electricity market has been to achieve efficiency through competition in supply and demand. A special feature of the electricity commodity is the reliance on a common network for transmission, where network

M. Bjørndal
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Østfold University College, 1757 Halden, Norway
e-mail: mette.bjorndal@nhh.no

K. Jörnsten
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: kurt.jornsten@nhh.no

L. Rud (✉)
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Institute for Research in Economics and Business Administration, Breiviksveien 40, 5045 Bergen, Norway
e-mail: linda.rud@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_16,
© Springer-Verlag Berlin Heidelberg 2010

constraints have important implications for optimal economic dispatch, due to the externalities created by the loop flow features of the network. Highly different market designs have been chosen for handling network constraints, with different implications for efficiency. Our objective is to combine several of the suggested approaches, and see whether it is possible to find a good approximation to optimal nodal prices by using system operator announced capacity charges.

In the proposed capacity charge approach (see also Bjørndal et al. 2000), capacity constraints are handled by issuing nodal capacity charges. Market clearing brings forth a single market price for energy which is common to the entire pool. The net nodal price equals the common market price less the nodal capacity charges. For positive (negative) capacity charges, the net nodal price is lower (higher) than the market price. Optimally set, capacity charges allow the market to reach optimal dispatch since net nodal prices then equal optimal nodal prices. Capacity charges are issued by the system operator, and are based partly on technical load factors, and partly on the shadow prices of congested lines. Implementation of the capacity charge approach can be on an ex post or an ex ante basis:

- If capacity charges are announced *ex post*, that is, after bidding and market clearing, the approach is merely a different representation of the nodal pricing method, now with the aggregate effect of capacity constraints in the grid priced explicitly in each node. With ex post announcement of capacity charges, the system operator has full information of shadow prices, and the calculation of capacity charges is straightforward.
- On an *ex ante* basis capacity charges are issued prior to bidding and market clearing, and are thus taken into consideration by the market participants when stating their supply and demand curves. In this case, now without full information, the shadow prices of congested lines have to be estimated. The implementation further depends upon the overall design of the electricity market. Within the framework of a *separate day-ahead scheduling market and a real-time balancing market*, the market participants state their bids after the announcement of capacity charges. Market clearing then brings forth a market equilibrium consistent with the estimated shadow prices of expected capacity constraints. The efficiency of the market equilibrium can be improved through iterative adjustment of the estimated capacity charges to reach an optimal and feasible market solution. Within the framework of a *pure real-time spot market*, the ex ante announcement of capacity charges allows participants to adjust demand and supply bids according to the anticipated capacity charge. These estimated capacity charges will, however, not be able to clear the market alone, and there is no room for a direct iterative process. This requires using, for example, ex post nodal prices together with the pre-announced capacity charges.

The capacity charge approach offers several advantages. A main issue is the role of capacity charges as important market signals for demand and supply, signaling geographical differences in the nodal cost of aggregate network constraints. Compared to approaches such as zonal pricing, the method also incorporates the advantages of nodal pricing, as capacity charges may be nodally differentiated. Further, as capacity

charges apply to all contracts of physical delivery, the method enables spot and bilateral contracts to coexist. Also, as capacity charges are issued by the system operator, and are based on technical information and estimates of shadow prices, this may enable a clearer distinction between the role of the exchange and the system operator, and might facilitate coordination through market coupling in areas where there exist multiple exchanges. As for implementing ex ante announced capacity charges in the case of the pure real-time spot market, the use of pre-announced capacity charges may be a source of enhancing market efficiency. In a market which is cleared basically at real-time only, we note that the producer or consumer has to be able to respond instantly to prices in order to submit price-elastic bids. If not, only price-inelastic bids can be submitted. The pre-announcement of capacity charges will enable this group to adjust plans according to the signals of expected congestion cost conveyed by the capacity charges.

The rest of the paper is organized as follows. Section 2 discusses the approach in relation to other methods for handling congestion in electricity markets. Section 3 presents the foundation for the capacity charge approach, and shows its relation to optimal dispatch by nodal prices, using a model with a "DC"-approximated network. Section 4 illustrates the approach of optimal capacity charges in a six-node example. Section 5 discusses iterative approaches for implementing the capacity charge method, illustrated with a standard gradient method. Section 6 discusses a heuristic approach for obtaining feasible flows. Section 7 concludes the paper, and future research is discussed.

2 Literature Review

The concept of nodal prices is discussed by Schweppe et al. (1988). Optimal nodal prices are produced by the solution of the welfare maximization problem as the dual prices of the power flow equations, and are interpreted as the value of power in each node (cf. Wu et al. (1996)). A mechanism that enforces optimal nodal prices to which generators and consumers adapt, will then contribute to social optimum in the short run. Wu et al. (1996), however, point to several counter-intuitive and possibly troublesome characteristics of implementing the nodal pricing approach. For instance, for the system operator to calculate the optimal economic dispatch and implement it, suppliers and consumers must truthfully reveal cost and demand functions, and they may not be willing to give away such strategic information.

On the other hand, the price system suggested by Chao and Peck (1996) represents a system for "explicit congestion pricing," where, instead of providing locational electricity prices as nodal prices do, the use of scarce transmission resources is explicitly priced. This is accomplished by the design of a trading rule based on load factors or distribution factors, which specify the transmission capacity rights that traders must acquire in order to complete an electricity transaction. In optimum, Chao–Peck prices are consistent with optimal nodal prices, and in accordance with the shadow prices of the transmission constraints of the optimal dispatch problem. A slight modification of Chao–Peck prices is suggested by Stoft (1998),

where a "hub" price is determined by allowing electricity bids at a given node (or "hub") in the network. Both mechanisms rest upon the existence of a market that brings forth the prices of transmission rights on the links. The number of prices these systems have to derive is usually far less than the number of nodes in the network.

The coordinated multilateral trade model suggested by Wu and Varaiya (1995) intends to attain optimal dispatch without requiring the system operator to collect private information, that is, supply and demand curve bids. Instead, brokers carry out profitable multilateral trades under feasibility constraints. Central coordination is achieved through an iterative process, where loading vectors are announced by the system operator, and where brokers using this information must evaluate the feasibility of the trades in question. Consequently, the decision mechanisms regarding economics and the feasibility of system operation are separated. Economic decisions are carried out by private multilateral trades among generators and consumers, while the function of ensuring feasibility is coordinated through the system operator who provides publicly accessible data, based upon which generators and consumers can determine profitable trades that meet the secure transmission loading limits.

In relation to the optimal dispatch problem, the coordination models can be interpreted as different relaxation schemes, with competitive players in generation and consumption, and where the system operator solves different sub-problems and information is exchanged back and forth. The decompositions corresponding to nodal pricing and Chao–Peck pricing are price-driven. In the case of nodal prices, the system operator hands out the optimal nodal prices obtained after solving the optimal dispatch problem, and optimal dispatch is achieved as producers and consumers adapt to their local prices. For Chao–Peck prices, a market is supposed to bring forth the competitive prices of transmission rights, while the system operator provides information on how trades affect every single link. When traders adapt to the transmission charges of the links imposed by the prices of the transmission rights, the overall problem is solved. The coordinated multilateral trade model can be interpreted as a Benders' decomposition method, where the market players maximize net profit, and quantities are communicated to the system operator, which in turn checks feasibility and generates constraints. The new constraints must then be taken into consideration when additional trades are placed, and the process continues. Due to the complexity of electric networks, each method has its limitations in practical use.

In this paper, we combine several of the above approaches. Our objective is to find good approximations of the optimal nodal prices based on an uncongested system price and the loading vectors of congested lines. This approach may be interpreted as a Chao–Peck pricing approach including a "hub," as suggested by Stoft (1998), where we estimate/guess the shadow prices of congested lines. Our approach is also similar to the coordinated multilateral trade model of Wu and Varaiya (1995) in that we need not rely on the disclosure of private information. Compared to Wu and Varaiya (1995), instead of announcing the constraints through the publication of the loading vector, the grid operator announces a set of nodal capacity charges that is based on an estimate/guess of the shadow price of the constraint in question, and the loading vector. The approach is also similar to that of Glavitch and Alvarado (1997) who use market information to estimate cost parameters.

3 The Capacity Charge Approach

The optimal market equilibrium is the market solution which gives the maximum social surplus attainable within the constraints of the system. In this section, we compare the market equilibrium of the capacity charge approach with this optimal economic dispatch.

Consider an electricity market where supply and demand are located in n nodes which are interconnected by a constrained transmission grid. Demand in node i, q_i^d, depends upon the market price p_i in the node. The demand curve is specified as the general function $p_i^d(q_i^d)$ which is assumed to be non-increasing in q_i^d. Likewise, supply in node i, q_i^s, also depends on the market price p_i of the node, and the supply curve is specified as the general non-decreasing function $p_i^s(q_i^s)$.

The social surplus, Π_{ss}, is defined as total willingness to pay, less total cost of production, as given in (1). The total willingness to pay is given by the area under the demand curve, while the total cost of production is the area under the supply curve.

$$\Pi_{ss} \equiv \sum\nolimits_{i=1}^{n} \left[\int_0^{q_i^d} p_i^d(q)dq - \int_0^{q_i^s} p_i^s(q)dq \right] \tag{1}$$

Assuming locational marginal prices in the various nodes, social surplus may be decomposed into demand surplus, supply surplus and grid revenue, the latter due to congestion. These are shown respectively as the three terms of (1′).

$$\begin{aligned}\Pi_{ss} = &\sum\nolimits_{i=1}^{n} \left[\int_0^{q_i^d} \left[p_i^d(q) - p_i^d(q_i^d) \right] dq \right] + \sum\nolimits_{i=1}^{n} \left[\int_0^{q_i^s} \left[p_i^s(q_i^s) - p_i^s(q) \right] dq \right] \\ &+ \sum\nolimits_{i=1}^{n} \left[p_i^d(q_i^d) q_i^d - p_i^s(q_i^s) q_i^s \right] \end{aligned} \tag{1'}$$

In general, the transmission grid consists of several lines that connect the nodes. To illustrate the nodal capacity charge approach, we consider real power using the lossless linear "DC" approximation of power flow equations, with reactance equal to 1 on every link.[1] Each line ij of the network is defined by the two nodes i and j which it interconnects. Let q_{ij} be the flow along line ij. If positive, the flow is in the direction from node i to node j. If negative, the flow is in the direction from node j to node i. Under the lossless "DC" approximation we have $q_{ij} = -q_{ji}$. The net injection in node i is defined by:

$$q_i \equiv q_i^s - q_i^d \tag{2}$$

If positive, there is a net injection. If negative, there is a net withdrawal.

[1] The "DC" approximation is the customary approximation used in the literature when dealing with the management of transmission constraints. Under these assumptions, and with well-behaved cost and benefit functions, the optimal dispatch problem is convex. For the specifics of the "DC" approximation, see for instance Wu and Varaiya (1995), Chao and Peck (1996), or Wu et al. (1996). In the "DC" approximation both losses and reactive power are left out.

The power flow on each line is determined by Kirchhoff's junction rule, Kirchhoff's loop rule, and the Law of conservation of energy, given by (3)–(5). Kirchhoff's junction rule (3) states that the current flowing into any node is equal to the current flowing out of it. There are n nodes, and there are $n-1$ independent equations.

$$q_i = \sum\nolimits_{j \neq i} q_{ij} \quad i = 1, \ldots, n-1 \tag{3}$$

Equation (4) follows from Kirchhoff's loop rule, that states that the algebraic sum of the potential differences across all components around any loop is zero. The number of independent loops is given by $m-n+1$, where m is the number of lines in the grid. $(\mathbf{L}) = (L_1, \ldots, L_{m-n+1})$ is the set of independent loops[2] and L_ℓ is the set of directed arcs ij in a path going through loop ℓ.

$$\sum\nolimits_{ij \in L_\ell} q_{ij} = 0 \quad \ell = 1, \ldots, m-n+1 \tag{4}$$

The law of conservation of energy (5) states that, in the absence of losses, total generation equals total consumption.

$$\sum\nolimits_i q_i = 0 \tag{5}$$

For a given network and load, the power flows may be represented by load factors. Each load factor β_{ij}^{lm} shows the fraction of an injection in node l with withdrawal in node m that flows along line ij. Note that $\beta_{ji}^{lm} = -\beta_{ij}^{lm}$ and $\beta_{ij}^{ml} = -\beta_{ij}^{lm}$. Under the "DC" approximation the load factors are constants, that is, they are independent of load.[3] By introducing a reference point r, the load factors may be represented by a loading vector $\boldsymbol{\beta}_{ij}(r) \equiv (\beta_{ij}^{1r} \beta_{ij}^{2r} \ldots \beta_{ij}^{nr}) \equiv (\beta_{ij}^{1} \beta_{ij}^{2} \ldots \beta_{ij}^{n})$ for each link ij. Element k of loading vector $\boldsymbol{\beta}_{ij}(r)$ shows the flow along line ij if 1 MW is injected into node k and withdrawn at the reference point r. A trade between node l and m, may be viewed as a combined trade between nodes l and r, and nodes r and m. Thus, we have that $\beta_{ij}^{lm} = \beta_{ij}^{lr} + \beta_{ij}^{rm} = \beta_{ij}^{lr} - \beta_{ij}^{mr} = \beta_{ij}^{l} - \beta_{ij}^{m}$.

Considering any line kl with net injections q_i given, the line flow along line kl can be expressed by load factors as:

$$q_{kl} = \sum\nolimits_i \beta_{kl}^i q_i \tag{6}$$

Capacity constraints $CAP_{ij} \geq 0$ and $CAP_{ji} \geq 0$ on line ij require that $q_{ij} \leq CAP_{ij}$ and $q_{ji} \leq CAP_{ji}$. The capacity constraints may thus be stated as:

$$\sum\nolimits_i \beta_{kl}^i q_i \leq CAP_{kl} \quad k = 1, \ldots, n, \quad l = 1, \ldots, n, \quad k \neq l \tag{7}$$

[2] See Dolan and Aldous (1993).

[3] In general AC systems the load factors depend on the distribution of loads over the network. Our method applies also for general AC systems, however, requiring recalculations of the load factors according to the load.

Note that if there is no direct link between nodes i and j, we have $\beta_{ij}^{lm} = \beta_{ji}^{lm} = 0$, and $CAP_{ij} = CAP_{ji} = 0$.

Under the "DC" approximation, with appropriate objective functions, for instance quadratic cost and benefit functions, the optimal economic dispatch is given by the following convex optimization problem:

$$\begin{aligned}
\textit{maximize} \qquad & \Pi_{ss} \equiv \sum_{i=1}^{n} \left[\int_0^{q_i^d} p_i^d(q)dq - \int_0^{q_i^s} p_i^s(q)dq \right] \\
\textit{subject to} \qquad & q_i = q_i^s - q_i^d \quad i = 1, \ldots, n \\
& q_i = \sum_{i \neq j} q_{ij} \quad i = 1, \ldots, n-1 \\
& \sum_{ij \in L_\ell} q_{ij} = 0 \qquad \ell = 1, \ldots, m-n+1 \\
& \sum_i q_i = 0 \\
& \sum_i \beta_{kl}^i q_i \leq CAP_{kl} \qquad k = 1, \ldots, n,\ l = 1, \ldots, n,\ k \neq l
\end{aligned} \tag{8}$$

In the unconstrained case, where neither of the capacity constraints of (8) are binding, there will be a uniform price in the market. For the capacity constrained case, where at least one capacity constraint is binding, nodal prices will differ and may be different for all nodes.[4] If the constraint $q_{kl} \leq CAP_{kl}$ is binding, we have $q_{kl} \geq CAP_{kl}$ and thus $q_{kl} \geq 0$. As $q_{lk} = -q_{kl}$, the corresponding constraint $q_{lk} \leq CAP_{lk}$ is not binding. Define the shadow prices of (7) as $\mu_{kl} \geq 0$. Thus, if $\mu_{kl} > 0$, we have $\mu_{lk} = 0$, and vice versa.

Under the capacity charge approach, we assume that the system operator first provides nodal capacity charges, cc_i. On receiving this information, the participants determine supply and demand bids. Market clearing results in an equilibrium energy price, p, which is common to the entire pool. The capacity charges may be positive or negative. A positive capacity charge, $cc_i > 0$, is defined as the amount cc_i that the suppliers in the node pay per unit supplied, or equivalently the amount cc_i that the consumers receive per unit consumed. For a negative charge, consumers pay, while producers are compensated. The net nodal price thus equals $p_i = p - cc_i$.

Proposition 1. *The market equilibrium of the capacity charge approach is in accordance with optimal economic dispatch when capacity charges are optimally defined.*

Proof. If we relax the capacity constraints in (8), we obtain the Lagrangian function:

$$\begin{aligned}
\Lambda(\boldsymbol{\mu}) = & \sum_i \left[\int_0^{q_i^d} p_i^d(q)dq - \int_0^{q_i^s} p_i^s(q)dq \right] \\
& + \sum_k \sum_l \mu_{kl} \left[CAP_{kl} - \sum_i \beta_{kl}^i (q_i^s - q_i^d) \right]
\end{aligned} \tag{9}$$

[4] Refer to Wu et al. (1996) for the characteristics of optimal nodal prices.

For a given vector $\boldsymbol{\mu}$ consisting of shadow prices for all lines of the network, the relaxed problem $h(\boldsymbol{\mu}) = \{\max \Lambda(\boldsymbol{\mu})\ \text{s.t.(2)-(5)}\}$ provides an upper bound on the objective function value of (8). This follows from weak duality. Because of strong duality, solving the dual problem $\min_{\boldsymbol{\mu}} h(\boldsymbol{\mu})$ also provides the solution to our original problem (8). Considering the objective function of the dual problem and rearranging terms, we get:

$$\begin{aligned}\Lambda(\boldsymbol{\mu}) &= \sum_i \left[\int_0^{q_i^d} p_i^d(q)dq + \sum_k \sum_l \mu_{kl}\beta_{kl}^i q_i^d \right] \\ &\quad - \sum_i \left[\int_0^{q_i^s} p_i^s(q)dq + \sum_k \sum_l \mu_{kl}\beta_{kl}^i q_i^s \right] + \sum_k \sum_l \mu_{kl} CAP_{kl} \\ &= \sum_i \int_0^{q_i^d} \hat{p}_i^d(q)dq - \sum_i \int_0^{q_i^s} \hat{p}_i^s(q)dq + \sum_k \sum_l \mu_{kl} CAP_{kl} \end{aligned} \tag{10}$$

The rearranged Lagrangian function is quite similar to the original (1), however, with two alterations. First, the original supply and demand functions are perturbed, and have been shifted by the term $\sum_k \sum_l \mu_{kl}\beta_{kl}^i$, as shown in (11).

$$\begin{aligned}\hat{p}_i^d(q_i^d) &= p_i^d(q_i^d) + \sum_k \sum_l \mu_{kl}\beta_{kl}^i = p_i^d(q_i^d) + cc_i \\ \hat{p}_i^s(q_i^s) &= p_i^s(q_i^s) + \sum_k \sum_l \mu_{kl}\beta_{kl}^i = p_i^s(q_i^s) + cc_i \end{aligned} \tag{11}$$

This perturbation is equivalent to the shift in supply and demand curves resulting from the capacity charge approach, where suppliers and consumers in node i face a capacity charge

$$cc_i = \sum_k \sum_l \mu_{kl}\beta_{kl}^i \tag{12}$$

Secondly, we have the addition of the last term $\sum_k \sum_l \mu_{kl} CAP_{kl}$. For a given shadow price vector μ, this term is a constant. For optimal shadow prices, μ_{kl}^*, the term is equal to the Merchandising surplus (cf. Wu et al. (1996)), which is equivalent to our definition of grid revenue in (1′). Thus, social optimum is achieved by the system operator issuing optimal capacity charges $cc_i^* = \sum_k \sum_l \mu_{kl}^* \beta_{kl}^i$, and subsequently solving the unconstrained optimal dispatch problem by clearing the market according to the perturbed supply and demand functions of (11). ■

In our approach, capacity constraints are managed by means of nodal capacity charges, which cause shifts in the supply and demand curves. Thus, constraints are implicitly taken care of, and market equilibrium results from clearing the market on a common energy price, p, that is, the system price. In optimum, the net prices of each node, $p_i = p - cc_i$, are equivalent to the optimal nodal prices of the nodal pricing approach. Note, however, that although optimal capacity charges cc_i are given by (12), they are not uniquely defined, but are associated with the load factors. When using load factors associated with a reference point, both the level of the capacity charges, and the system price are affected by the chosen reference

point. Using optimal shadow prices will however always ensure the same optimal net nodal prices, regardless of which reference point is chosen. If market participants optimally adapt to the net price, that is, the market energy price corrected by the capacity charge, market equilibrium is not affected by the choice of reference point.

4 A Numerical Example

To illustrate the nodal capacity charge approach, we use an electricity market model with production and consumption located in six nodes. As a benchmark we show the outcome of the optimal unconstrained and constrained dispatch, the latter using nodal prices. With optimally defined capacity charges, we show that the outcome of the capacity charge approach is identical to that of nodal pricing.

4.1 Model and Parameters

In the example, we assume that generators have quadratic cost functions, with a profit function π^s of the general form $\pi^s = (p - cc_i)q_i^s - \frac{1}{2}c_i(q_i^s)^2$, which gives us linear supply curves. We also assume linear demand functions. Supply and demand curves, including capacity charges, are shown as:

$$p = c_i q_i^s + cc_i \tag{13}$$

$$p = a_i + cc_i - b_i q_i^d \tag{14}$$

where a_i, b_i, and c_i are positive parameters. The parameters of our numerical example are shown in Table 1.

Social surplus, decomposed into the surpluses of consumers and suppliers, and grid revenue due to congestion, is given by (15).

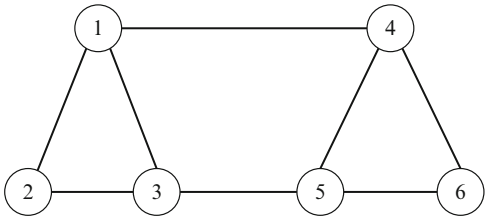
$$\Pi_{ss} \equiv \sum_{i=1}^{6} \tfrac{1}{2}(a_i - p + cc_i)q_i^d + \sum_{i=1}^{6} \tfrac{1}{2}(p - cc_i)q_i^s + \sum_{i=1}^{6} cc_i(q_i^s - q_i^d) \tag{15}$$

The network connecting the six nodes is shown in Fig. 1.

Table 1 Parameters

Node	a_i	b_i	c_i
1	20	0.05	0.2
2	20	0.05	0.1
3	30	0.10	0.7
4	20	0.05	0.2
5	30	0.10	0.7
6	30	0.10	0.1

Fig. 1 Network



We apply the lossless linear "DC" approximation of the power flow equations, with reactance equal to 1 on every link. For given net injections q_i, power flows are determined according to (3)–(5). Thus, the node rule equations follow in (3′), the loop rule equations[5] in (4′) and the requirement of energy balance in (5′).

$$\begin{aligned} q_1 &= q_{12} + q_{13} + q_{14} \\ q_2 &= -q_{12} + q_{23} \\ q_3 &= -q_{23} - q_{13} + q_{35} \\ q_4 &= -q_{14} + q_{45} + q_{46} \\ q_5 &= -q_{35} - q_{45} + q_{56} \end{aligned} \tag{3′}$$

$$\begin{aligned} q_{13} &= q_{12} + q_{23} \\ q_{13} &= q_{14} + q_{45} - q_{35} \\ q_{13} &= q_{14} + q_{46} - q_{56} - q_{35} \end{aligned} \tag{4′}$$

$$q_1 + q_2 + q_3 + q_4 + q_5 + q_6 = 0 \tag{5′}$$

The load factors β_{ij}^{lm} show the power flow on line ij following from an injection of one unit in node l and withdrawing it in node m. By solving (3′)–(5′) with $q_l = 1$ and $q_m = -1$, we can compute the load factors β_{ij}^{lm} for all links. The load factors for our example network are shown in Table 2.

For example, a trade consisting of injecting 1 MW in node 1 and withdrawing it in node 3, gives a flow over line 12 equal to $\frac{4}{15}$ in direction from 1 to 2. Likewise, injecting 1 MW in node 2 and withdrawing it in node 3, results in a flow over line 12 of $-\frac{11}{30}$, that is, a flow of $\frac{11}{30}$ from 2 to 1. If combined, the flows over line 12 from the two trades partly cancel, resulting in a net flow over line 12 equal to $q_{12} = \frac{4}{15} + (-\frac{11}{30}) = \frac{1}{10}$.

Alternatively, by introducing a reference point r for withdrawals, the load factors may be represented by the loading vectors $\boldsymbol{\beta}_{ij}(r)$ for each link ij. The load factors using reference point 3 are for example given by β_{ij}^{l3} derived from the five columns for trades with node 3 in Table 2, (columns 13, 23, 34, 35, and 36), and noting that $\beta_{ij}^{ml} = -\beta_{ij}^{lm}$ and that $\beta_{ij}^{33} = 0$ by definition. This loading vector shows line flows for trades from any injection point to the reference point only. Note that all

[5] The number of independent loops are $m - n + 1 = 8 - 6 + 1 = 3$.

Table 2 Load factors

		Trades (*lm*)														
		12	13	14	15	16	23	24	25	26	34	35	36	45	46	56
	12	19/30	4/15	1/10	1/6	2/15	−11/30	−8/15	−7/15	−1/2	−1/6	−1/10	−2/15	1/15	1/30	−1/30
	13	4/15	8/15	1/5	1/3	4/15	4/15	−1/15	1/15	0	−1/3	−1/5	−4/15	2/15	1/15	−1/15
	14	1/10	1/5	7/10	1/2	3/5	1/10	3/5	2/5	1/2	1/2	3/10	2/5	−1/5	−1/10	1/10
Lines	*23*	−11/30	4/15	1/10	1/6	2/15	19/30	7/15	8/15	1/2	−1/6	−1/10	−2/15	1/15	1/30	−1/30
(*ij*)	*35*	−1/10	−1/5	3/10	1/2	2/5	−1/10	2/5	3/5	1/2	1/2	7/10	3/5	1/5	1/10	−1/10
	45	1/15	2/15	−1/5	1/3	1/15	1/15	−4/15	4/15	0	−1/3	1/5	−1/15	8/15	4/15	−4/15
	46	1/30	1/15	−1/10	1/6	8/15	1/30	−2/15	2/15	1/2	−1/6	1/10	7/15	4/15	19/30	11/30
	56	−1/30	−1/15	1/10	−1/6	7/15	−1/30	2/15	−2/15	1/2	1/6	−1/10	8/15	−4/15	11/30	19/30

information of Table 2 is contained in these five columns. A general trade between, for example, node 1 and 6, may be viewed as a combined trade between node 1 and 3 and between 3 and 6. For example, the flow on line 12 resulting from this trade is $\beta_{12}^{16} = \beta_{12}^{13} + \beta_{12}^{36} = \frac{4}{15} + (-\frac{2}{15}) = \frac{2}{15}$.

4.2 Unconstrained Optimal Dispatch

Assuming no congestion in the network, optimal dispatch and maximum social surplus results from aggregating supply and demand curves, and clearing the market so that the prices of all regions are the same, that is, $p_1 = p_2 = p_3 = p_4 = p_5 = p_6$. Due to the absence of constraints, all resulting flows are feasible, and capacity charges and grid revenue due to congestion are thus equal to 0. In our example, the market price of energy in the scenario of zero capacity charges is 17.09. Table 3 shows the optimal unconstrained dispatch of our example. The table also displays total social surplus, 7552.33, and its allocation to production, consumption, and the grid.

4.3 Nodal Prices

Now, assume that the capacity of the lines are as shown in Table 4. In the example we assume that $CAP_{ij} = CAP_{ji}$[6]. These capacity constraints make the unconstrained optimal dispatch infeasible, as the constraints of lines 23, 35 and 45 are violated at this solution.

With the nodal pricing approach optimal nodal prices result from optimal dispatch. Table 5 displays optimal dispatch, nodal prices, and the allocation of social surplus. In optimal dispatch, we find that the capacity of lines 23, 45, and 46 are

Table 3 Unconstrained dispatch

Node	Price	Supply	Demand	Net Injection	Supply Surplus	Demand Surplus	Grid Revenue	Total Surplus	Line	Flow
									12	−34.40
1	17.09	85.47	58.14	27.33	730.43	84.51	0.00	814.93	13	43.99
2	17.09	170.93	58.14	112.79	1460.86	84.51	0.00	1545.36	14	17.73
3	17.09	24.42	129.07	−104.65	208.69	832.95	0.00	1041.64	23	78.39
4	17.09	85.47	58.14	27.33	730.43	84.51	0.00	814.93	35	17.73
5	17.09	24.42	129.07	−104.65	208.69	832.95	0.00	1041.64	45	43.99
6	17.09	170.93	129.07	41.86	1460.86	832.95	0.00	2293.81	46	1.07
Sum		561.63	561.63	0.00	4799.96	2752.37	0.00	7552.33	56	−42.93

[6] In reality, this may not be so, as the grid lines may be operated with different capacities depending on the direction of the flow over the interconnection. This is especially so if links are aggregated individual lines.

Table 4 Line capacities

Line	Capacity
12	60
13	60
14	60
23	60
35	10
45	30
46	8
56	60

binding. Line 35, although expected, is not constrained, while the flow direction of line 46 has changed and the flow limit is binding in optimal dispatch. The table also shows optimal shadow prices for each line. Since $\mu_{ji} = 0$ if $\mu_{ij} > 0$, we have displayed only one shadow price per line. If positive, it indicates that the constraint CAP_{ij} is binding. If negative, it indicates that the constraint CAP_{ji} is binding, where the absolute value of μ_{ij} is the shadow price of the constraint CAP_{ji}.

The nodal pricing approach follows from Schweppe et al. (1988), see also Hogan (1992). In order to implement a system of nodal prices, it is required that the system operator calculates optimal nodal prices on the basis of information of the network, supply, and demand.

4.4 Optimal Capacity Charges

Under the capacity charge approach a positive or negative capacity charge cc_i is issued to each node, while the market is cleared at a single equilibrium price p. If capacity charges are announced prior to market clearing, consumers and producers will take the announced capacity charge into account when deciding supply and demand bids. Market response to optimally defined capacity charges, will result in a feasible and optimal market equilibrium.

Table 6 shows the optimal capacity charges, when the system price is defined as the unconstrained energy price. Note that the net nodal prices, $p_i = p - cc_i$, equal the optimal nodal prices, and that the resulting market equilibrium and social surplus of the two methods coincide. Likewise, if capacity charges are announced *after* bidding and market clearing, we see that it is straightforward to represent nodal prices by a common market price and nodal capacity charges, so that $p_i = p - cc_i$, that is $cc_i = p - p_i$.

Optimal capacity charges are defined by (12), using the optimal shadow prices from Table 5, and load factors defined by the physical characteristics of the grid from Table 2. Load factors are defined relatively to the chosen reference point. The level of both the system price and the capacity charges depend on this chosen point of reference. The net nodal price, $p_i = p - cc_i$, as well as the nodal differences between both net prices and between capacity charges, however, are the same, regardless of the chosen point of reference. Table 7 shows examples of optimal sets of energy price and capacity charges, depending on the chosen reference point.

Table 5 Optimal dispatch: nodal prices

Node	Price	Supply	Demand	Net Injection	Supply Surplus	Demand Surplus	Grid Revenue	Total Surplus	Line	Flow	Shadow price
									12	−24.42	0.00
1	17.05	85.23	59.08	26.15	726.42	87.26	−445.78	813.68	13	35.58	0.00
2	16.15	161.47	77.05	84.42	1303.69	148.43	−1363.17	1452.11	14	14.99	0.00
3	18.71	26.73	112.89	−86.17	250.06	637.26	1612.20	887.32	23	60.00	3.46
4	16.28	81.40	74.39	7.01	662.62	138.36	−114.08	800.98	35	9.41	0.00
5	19.48	27.82	105.24	−77.41	270.95	553.74	1507.73	824.69	45	30.00	6.14
6	17.30	173.00	127.00	46.00	1496.45	806.45	−795.80	2302.90	46	−8.00	−1.16
Sum		555.66	555.66	0.00	4710.18	2371.50	401.11	7482.79	56	−38.00	0.00

Table 6 Optimal capacity charges

Node	Price	Capacity Charge	Supply	Demand	Net Injection	Supply Surplus	Demand Surplus	Grid Revenue	Total Surplus	Line	Flow
										12	−24.42
1	17.09	0.05	85.23	59.08	26.15	726.42	87.26	1.23	814.91	13	35.58
2	17.09	0.95	161.47	77.05	84.42	1303.69	148.43	79.83	1531.95	14	14.99
3	17.09	−1.62	26.73	112.89	−86.17	250.06	637.26	139.37	1026.69	23	60.00
4	17.09	0.81	81.40	74.39	7.01	662.62	138.36	5.69	806.68	35	9.41
5	17.09	−2.38	27.82	105.24	−77.41	270.95	553.74	184.50	1009.19	45	30.00
6	17.09	−0.21	173.00	127.00	46.00	1496.45	806.45	−9.52	2293.38	46	−8.00
Sum			555.66	555.66	0.00	4710.18	2371.50	401.11	7482.79	56	−38.00

As market participants face identical net nodal prices in all cases, their resulting supply and demand will be the same as in optimal dispatch, that is, as shown in Table 5. We find the same production, consumption, line flows, social surplus, and allocation of surplus, including identical grid revenues. Moreover, the grid revenue is equal to the merchandizing surplus under optimal nodal prices.

However, although all market participants are equally well off in all cases, we may find that their perception of the situations may differ. For the individual market participant, it may be difficult to see the relation between the market price and the capacity charges. A market participant facing a capacity charge would thus be likely to think of his burden due to the constraint as the capacity charge cc_i he faces, with a total burden of $cc_i q_i^s$ for suppliers and $-cc_i q_i^d$ for consumers. Table 8 displays the *perceived* burdens of the consumers and producers due to the transmission constraints.

For instance, considering the producer in node 6, the choice of node 5 as the reference node leads to a total payment of 376.50, whereas the choice of node 2 as the reference point, induces a total compensation of 199.41. However, since the net prices are identical in all situations, the surpluses of each participant will be the same as shown in Table 5. For example, the producer in node 6 has a supply surplus of 1496.45 in all cases.

5 An Iterative Adjustment Process

We found that optimal capacity charges and appropriate adjustment of bid curves lead to optimal dispatch. From (12) we see that the informational requirements for issuing optimal capacity charges are the loading vectors and the shadow prices of congested lines. Loading vectors are technical information, which we assume are readily available. Shadow prices are in principle found by solving the optimal dispatch problem, thus requiring that the system operator has information on cost and benefit functions. When capacity charges are issued prior to bidding and market clearing, shadow prices have to be estimated. Estimated shadow prices can be improved through an iterative process, making use of market responses to obtain

Table 7 Optimal capacity charges and energy prices

		Basis of determining capacity charges						
		Unconstrained price	Reference point 1	Reference point 2	Reference point 3	Reference point 4	Reference point 5	Reference point 6
Energy price		17.09	17.05	16.15	18.71	16.28	19.48	17.30
Nodal capacity charges	1	0.05	0.00	−0.90	1.66	−0.77	2.43	0.25
	2	0.95	0.90	0.00	2.56	0.13	3.33	1.15
	3	−1.62	−1.66	−2.56	0.00	−2.43	0.77	−1.41
	4	0.81	0.77	−0.13	2.43	0.00	3.20	1.02
	5	−2.38	−2.43	−3.33	−0.77	−3.20	0.00	−2.18
	6	−0.21	−0.25	−1.15	1.41	−1.02	2.18	0.00

Table 8 Perceived burdens of the constraint

	Basis of determining capacity charges						
	Unconstrained price	Reference point 1	Reference point 2	Reference point 3	Reference point 4	Reference point 5	Reference point 6
Producer 1	4.00	0.00	−76.60	141.86	−65.27	207.13	21.64
Consumer 1	−2.77	0.00	53.09	−98.33	45.24	−143.58	−15.00
Producer 2	152.70	145.12	0.00	413.88	21.46	537.54	186.12
Consumer 2	−72.87	−69.25	0.00	−197.50	−10.24	−256.50	−88.81
Producer 3	−43.23	−44.49	−68.51	0.00	−64.96	20.47	−37.70
Consumer 3	182.61	187.91	289.37	0.00	274.36	−86.45	159.24
Producer 4	66.16	62.34	−10.82	197.83	0.00	260.16	83.01
Consumer 4	−60.46	−56.97	9.89	−180.80	0.00	−237.77	−75.86
Producer 5	−66.31	−67.62	−92.62	−21.31	−88.92	0.00	−60.55
Consumer 5	250.81	255.75	350.33	80.59	336.34	0.00	229.03
Producer 6	−35.81	−43.93	−199.41	244.02	−176.41	376.50	0.00
Consumer 6	26.29	32.25	146.39	−179.14	129.50	−276.39	0.00
Sum	401.11	401.11	401.11	401.11	401.11	401.11	401.11

good estimates of the shadow prices. Such an iterative approach is similar to that of Wu and Varaiya (1995), however, while they use feasibility and clever market agents (brokers) to arrange multilateral trades, we use prices and market response to coordinate the nodal markets.

The problem of the system operator in this case is to state capacity charges, based on estimates of the shadow prices of the congested lines, and improved upon through an iterative process. The iterative approach may be interpreted and implemented in a direct or indirect manner. The direct approach involves a series of actual iterations in clearing the market, where participants after each market clearing receive adjusted capacity charges to which they respond with adjusted supply and demand curve bids. The final market price and capacity charges will be those of the last iteration. While this direct approach contributes to ensuring "correct" prices for each point in time, the transaction costs of several iterations for each market clearing could be quite large. Alternatively, the iterative approach may be implemented indirectly. On observing a pattern of congestion similar to earlier periods, the iteration comes about when the system operator uses information on earlier market responses to improve estimates. This may be a more cost efficient method, and justifiable if congestion situations last for a period of time. In this case, market responses to earlier capacity charges may be used to obtain better estimates. It is also possible to start with an estimate based on information obtained from earlier estimates and market observations, and improve capacity charges through a few iterations.

We illustrate the iterative process by using a simple updating procedure in our example, assuming the direct interpretation, or alternatively, identical market conditions in consecutive time slots. In the example we start with capacity charges equal to zero, implying that no lines are congested. This results in the unconstrained dispatch solution, which is not feasible. Alternatively, starting points may be based on forecasts of congested lines and shadow prices.

In each iteration, shadow prices, and consequently capacity charges, are updated, to relieve the congested lines. The objective here is to illustrate the approach, rather than finding the most efficient updating rule. There is a vast literature on algorithms for updating (see for example Minoux 1986). For illustration, we have employed a standard gradient method, where the shadow prices are updated on the general form:

$$\mu_{ij}^{t+1} = \mu_{ij}^{t} + \lambda_{ij}^{t} \frac{\gamma_{ij}^{t}}{\left\| \gamma_{ij}^{t} \right\|} \tag{16}$$

where μ_{ij}^{t} is the estimated shadow price of line ij at iteration t, γ_{ij}^{t} is a gradient of the objective function in (9) valued at iteration t, $\left\| \gamma_{ij}^{t} \right\|$ is a normalization of the gradient, and λ_{ij}^{t} is the step chosen at time t. In the example, we have defined the terms as follows:

$$\gamma_{ij}^{t} = \frac{\partial \Lambda^{t}}{\partial \mu_{ij}^{t}} = CAP_{ij} - \sum_{k} \beta_{ij}^{k} q_{k}^{t} = CAP_{ij} - q_{ij}^{t} \tag{17}$$

$$\left\| \gamma_{ij}^{t} \right\| = CAP_{ij} \tag{18}$$

where we see that γ_{ij}^{t} is the under- or over-utilization of the line, and by normalizing by CAP_{ij}, we have the relative under- or over-utilization of the line.

If $CAP_{ij} - q_{ij} < 0$, the line is congested, requiring the shadow price estimate of line ij to be raised. If $CAP_{ij} - q_{ij} > 0$, the line is not congested and any shadow price estimate for the line has to be driven towards 0. Thus the step λ_{ij}^{t} must be negative. The size of the step λ_{ij}^{t} determines the speed of change. In our example relatively small steps induce a slow convergence towards the optimal value, while larger, but still moderate steps give a faster convergence. However, steps which are large relative to the congestion of the line may cause an oscillation of capacity utilization and shadow price around the optimal values. A definition of step size dependent on the degree of capacity utilization, for example as shown in (19), may give a faster convergence when over-utilization is high, while reducing oscillation around the optimal value.

$$\lambda_{ij}^{t} = \begin{cases} -1 & \text{for} \quad \dfrac{\left| CAP_{ij} - q_{ij}^{t} \right|}{CAP_{ij}} > 0.1 \\ -0.1 & \text{for} \quad \dfrac{\left| CAP_{ij} - q_{ij}^{t} \right|}{CAP_{ij}} \leq 0.1 \end{cases} \tag{19}$$

Figure 2 shows the resulting development of social surplus in 25 iterations defined by (16)–(19) and with non-negativity constraints on the shadow prices. Starting from the value connected to the infeasible unconstrained case, social surplus evolves towards the level of the optimal case, however, oscillating due to our rather crude definition of the iteration process.

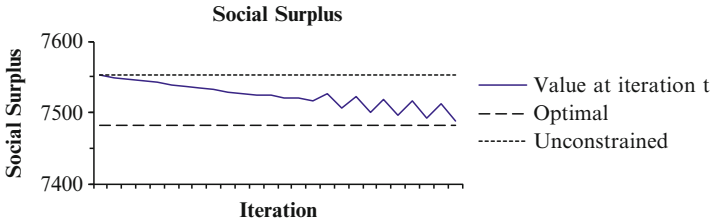


Fig. 2 Iterations: social surplus

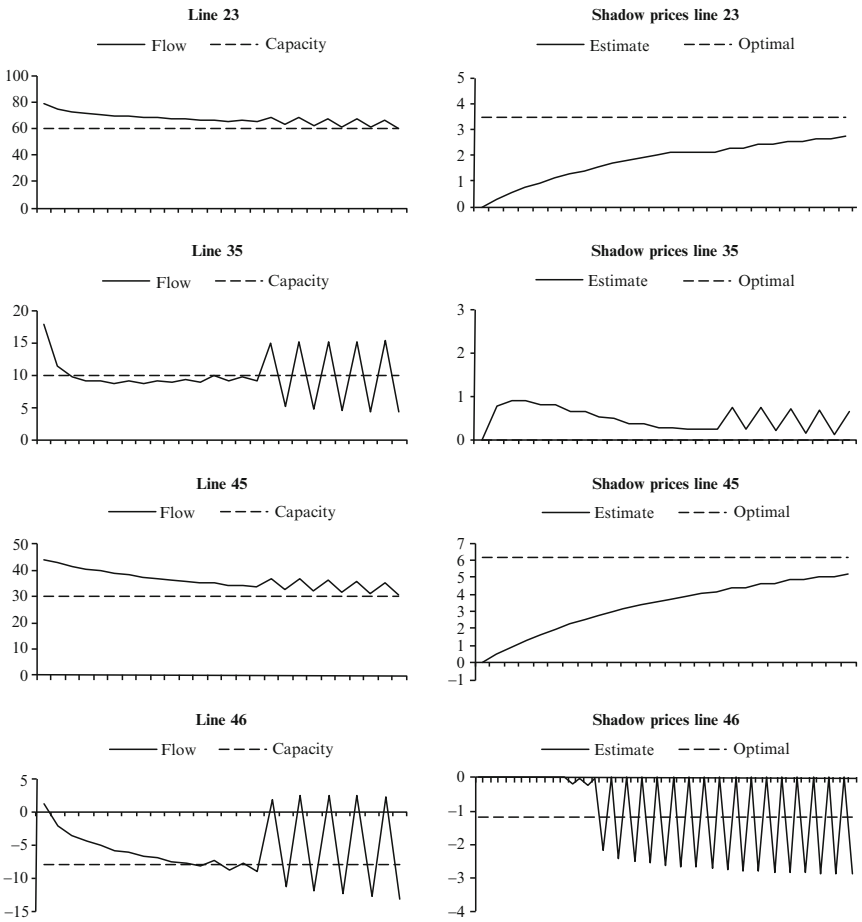


Fig. 3 Iterations: capacity utilization and shadow prices

The corresponding developments of capacity utilization and shadow prices are displayed in Fig. 3. The left hand side displays the capacity and capacity utilization of the lines that either are constrained starting with zero capacity charges (that is, at

the system price of unconstrained dispatch), and/or that are constrained in optimal dispatch. As in the above tables, the capacity and flow of line ij is displayed as a positive number when the flow is in the direction i to j, and as a negative number when the flow is in the direction j to i. The right hand side shows the estimated shadow prices, where we equivalently have determined $\mu_{ij} > 0$ as the shadow price for the capacity in the direction of i to j, while $|\mu_{ij}|$ is the shadow price of capacity in the direct from j to i if $\mu_{ij} < 0$.

In the unconstrained market solution, we find that the implied flows over lines 23, 35, and 45 exceed capacity. This indicates a positive shadow price for each of these lines, and the adjustment rule induces a rise in the shadow price estimates. Nodal capacity charges are issued based on these estimates, and the result in the second round of market clearing is a reduction of the actual flow over these lines. Further iterations gradually increase the shadow price estimates of the lines, thus reinforcing the reduction of the flow over the capacitated lines towards a feasible solution.

It should be noted that the flows over all lines change as a result of the capacity charges and the corresponding market equilibria:

- Line 35 is initially infeasible in the unconstrained solution, and the shadow price of this line is estimated to be positive. However, though initially infeasible, changes in the flows of other lines due to capacity charges, actually induce a lower flow over line 35, inducing a flow below capacity in optimum.
- Lines 12, 13, 14, and 56 are initially feasible. New market solutions as a result of the capacity charges cause changes in lines 12, 13, 14, and 56 with flows still below the line capacity.
- Line 46 starts with an initial feasible flow of 1.07 in the direction from node 4 to node 6. Changes elsewhere in the network give a reduced flow, and subsequently a change in the flow direction, from node 6 to node 4, as illustrated in the figure by negative numbers. Further changes in the network flows implicate a non-feasible flow from 6 to 4 and thus a positive shadow price estimate for μ_{64}.

The iterations in Fig. 3 illustrate the use of the gradient for updating shadow price estimates. In our example, we find that with a constant small step, for example, 0.1, the line flow is driven asymptotically towards the capacity, albeit necessitating a large number of iterations. A higher step size will speed the process when the shadow price is far from the optimal value, but results in an oscillation around the optimal value when coming near optimum, as shown in the example. The engineering of a more efficient algorithm will reduce the number of iterations called for. However, taking into account the costs of iterations, in order to obtain feasible flows within a small number of iterations, the adjustment procedure may have to be combined with some other mechanism as, for example, curtailment or counter trading. In the next section we attempt to speed up the convergence by a simple heuristic procedure.

6 A Heuristic Procedure for Faster Convergence

Above we have illustrated how iterations based on a simple standard gradient method can bring the market solution towards optimal dispatch, and reduce line flows of constrained lines towards capacity limits. We see that this procedure may require a rather large number of iterations to reach the optimal solution. If the cost of further iterations exceeds the gain in social surplus, it may be optimal to terminate the iterative procedure before reaching the optimal solution. However, note that the illustrated procedure is an upper bounding procedure, where the line flows of constrained lines are driven to the capacity limit from above. By prior termination of the iteration procedure, the resulting flow will not be feasible. To find a feasible flow, load or injections may be curtailed. A problem is, however, to curtail such that the resulting market equilibria, that is, prices and quantities, are consistent with bid functions. An infeasible flow can also be corrected through a secondary market, for instance for counter trading. An alternative or supplement to such cut-off mechanisms is to "force" the iteration itself to reach a feasible solution.

In this section we discuss a heuristic approach for finding a feasible flow that is also a market equilibrium, and as we will see, brings us near the optimal solution. The proposed heuristic procedure is based on Everett (1963).

Let us first summarize the problem. Focusing on the capacity constraints, the optimal dispatch problem of (8) may be formulated as follows:

$$\begin{array}{ll} \textit{maximize} & \Pi_{ss}(\mathbf{q}) \\ \textit{subject to} & q_{kl}(\mathbf{q}) \leq CAP_{kl} \qquad \forall\, kl \end{array} \tag{8'}$$

where $\mathbf{q} = (q_1^s, \ldots, q_n^s, q_1^d, \ldots, q_n^d)$ is the vector of production and consumption in each node, and $q_{kl}(\mathbf{q})$ is the flow on line kl. The shadow price vector $\boldsymbol{\mu}$ gives the shadow prices for all lines in both directions. The Lagrangian function is:

$$\Lambda(\boldsymbol{\mu}) = \Pi_{ss}(\mathbf{q}) + \sum_k \sum_l \mu_{kl} \left[CAP_{kl} - q_{kl}(\mathbf{q}) \right] \tag{9'}$$

In this setting Everett's theorem can be stated as follows:

1. Choose an arbitrary vector $\boldsymbol{\mu}$ of non-negative shadow prices for all lines.
2. Find a solution $\mathbf{q}^*$ which maximizes the unconstrained Lagrangian function $\Lambda(\boldsymbol{\mu})$.
3. Then, $\mathbf{q}^*$ is the solution to the constrained maximization problem with the original objective function in (8′), but with modified capacity limits CAP'_{kl}, given by $CAP'_{kl} = q_{kl}(\mathbf{q}^*)$.

The ex ante announcement of capacity charges is easily interpreted within this theorem. In order to set capacity charges according to (12), that is, $cc_i(\boldsymbol{\mu})$, the system operator has to choose a set of shadow prices $\boldsymbol{\mu}$. Market participants bid on the basis of the capacity charges, and by clearing the market on a single spot price, in effect we find the unconstrained maximum of the Lagrangian function $\Lambda(\boldsymbol{\mu})$.

The solution $\mathbf{q}^*$ is a true market equilibrium, and is both feasible and optimal with respect to the *modified* constraints $CAP'_{kl} = q_{kl}(\mathbf{q}^*)$. Our problem is to reach a feasible and optimal solution within the *real* constraints of the network, calling for an adjustment of shadow prices through an iterative process. Different sets of $\boldsymbol{\mu}$ lead to different flows. However, if the chosen $\boldsymbol{\mu}$ gives a feasible solution, and the flow actually equals the real constraint for binding constraints, the Everett theorem implies that this solution is the optimal solution to the original constrained problem.

Based on this insight, we have slightly modified the updating procedure. The main issue is to reach shadow prices (and thus capacity charges) that will give rise to a flow that matches the constraints. As a heuristic we attempt to speed movement towards a feasible solution by simply exaggerating the over-utilization of the grid by redefining the capacity size used in calculating both the gradient and the step. Starting with a market equilibrium that causes an infeasible flow over line ij, the updating procedure is still given by (16)–(19), with one alteration. We have substituted the real capacity CAP_{ij} with $CAP'_{ij} = \alpha_{ij} CAP_{ij}$, where $0 < \alpha_{ij} < 1$. The relative over-utilization of the line will then be exaggerated, the normalized gradient will be larger, and with the specified step of (19), the step might also be higher. This causes a faster adjustment towards shadow prices that give a feasible flow. Note that not all lines become feasible within the same iteration. Thus, even though a line becomes feasible for a market equilibrium, further iterations may be called for to obtain feasible flows on other lines. Subsequent iterations however slightly alter equilibrium supply and demand, causing changes in line flow over all lines. There will, thus, still occur changes in the line flow over lines that have become feasible. Therefore, the moment that the flow of a given line becomes feasible with respect to the real capacity, any further adjustments in shadow prices are made using the real capacity in (16)–(19).

In a market equilibrium, supply and demand must balance according to the net prices of the nodes, and produce a flow that is feasible. There exists infinitely many market equilibria, though they do not represent the optimal dispatch, cf. Wu et al. (1996). Market clearing ensures that supply and demand balances. By Everett's theorem we find that if the line flows defined by the market solution are feasible and equal to the capacity limit for binding constraints, the resulting market solution is indeed the optimal dispatch.

Figures 4 and 5 illustrate a simple version of the heuristic, where $\alpha_{ij} = 0.7$ is constant and equal for all lines. In Fig. 4, we see that social surplus more quickly advances to the optimal level of social surplus than without the heuristic (Fig. 2).

Figure 5 shows how the heuristic affects line flows and shadow price estimates. Even with this rather crude definition of the heuristic, using the same α_{ij} for all lines, we see that the line flows of the constrained lines become feasible in less iterations. Also note that the estimated shadow prices come close to the optimal shadow prices in relatively few iterations.

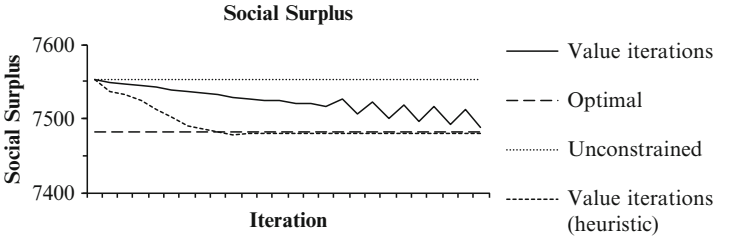


Fig. 4 Iterations: social surplus

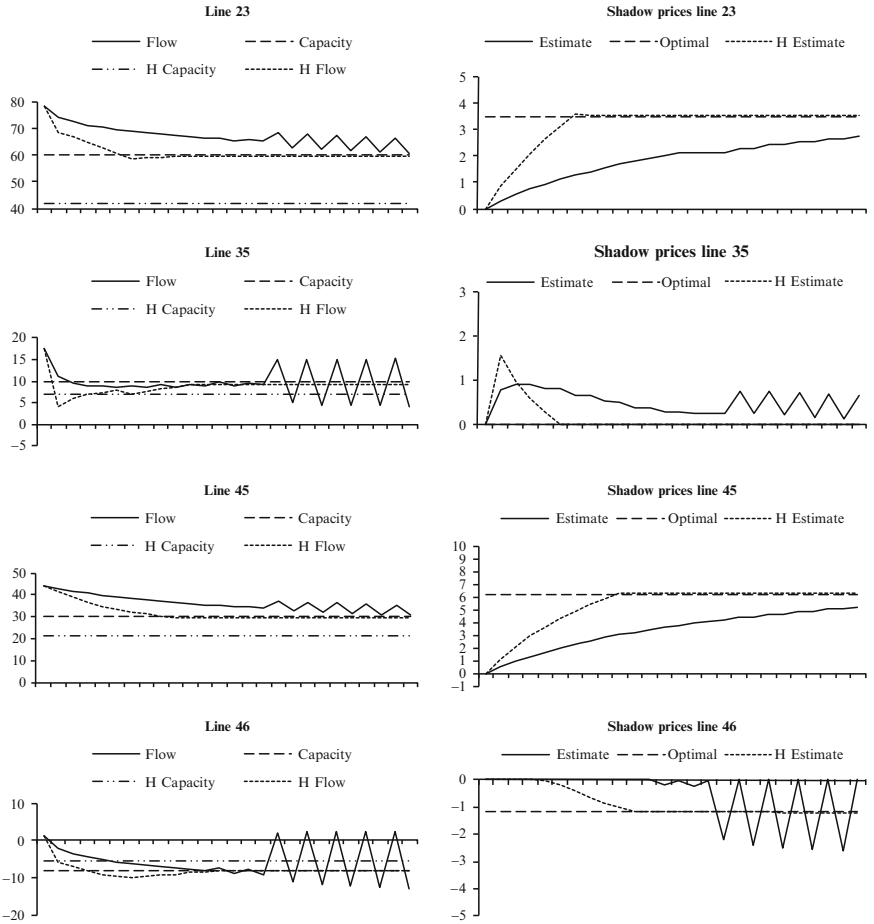


Fig. 5 Iterations: capacity utilization and shadow prices

7 Summary and Topics for Future Research

Electricity markets have been reorganized, introducing competition in supply and demand in order to achieve greater efficiency. Capacity constraints in electricity networks represent rather complex constraints on an efficient market solution, due to the externalities created by the loop flow. Several different market designs have been chosen for handling network constraints. Optimal nodal prices, as the solution of the welfare maximization problem, represent an efficient market equilibrium, where optimal nodal prices reflect both supply and demand conditions, together with the constraints of the network. In the capacity charge approach, we combine several of the suggested approaches. An objective is to find good approximations of the optimal nodal prices based on a competitive market price and nodal capacity charges issued by the system operator.

A main issue of the approach is the role of capacity charges as important market signals for demand and supply. The capacity charge approach is similar to a flowgate approach, but instead of signaling the congestion cost of individual lines, signals the nodal cost of aggregate congestion in the network. Compared to an approach such as zonal pricing, the method incorporates the advantages of nodal pricing. Capacity charges apply to all contracts of physical delivery, facilitating the co-existence of exchange traded and bilaterally traded contracts. The calculation of capacity charges is based on technical information, together with estimates of shadow prices. The capacity charges are issued by the system operator. This recognizes the coordination task performed by the system operator, as well as the system operator's access to important information on physical dispatch. At the same time, the approach enables a clearer distinction between the role of the exchange and the system operator, and might facilitate coordination in areas with separate exchanges. The efficiency of the approach is, however, contingent on the quality of shadow price estimates. We show that estimated shadow prices can be improved through a (direct or indirect) iterative process, making use of market responses to obtain good estimates of the shadow prices. A potential source of inefficiency is related to the system operator's incentives for stating capacity charges that boost their revenue. This is an issue for further investigation, and should be seen in connection with the regulation of the system operator.

In principle, capacity charges may be announced before or after market bidding. If capacity charges are announced after market clearing, the approach is in effect identical to the nodal pricing approach. When announced prior to bidding and market clearing, capacity charges give signals to the market as to expected congestion. Within a market design with a day-ahead scheduling market and a real-time balancing market, such as the Norwegian system, market participants act upon this information in stating their supply and demand bids. The efficiency of the market equilibrium can be improved through an iterative adjustment process to reach an optimal or near optimal feasible solution. Within a real-time spot market only (i.e., without the day-ahead market), the ex ante announcement of capacity charges signals expected congestion prices, and may induce a shift in supply and demand which otherwise would be inelastic with respect to the real-time spot price.

We propose that the capacity charge approach may be an effective procedure for managing transmission constraints in a competitive electricity market, for example in the Nordic scheduled day-ahead power market. The capacity charge approach bears a close relation to the system used today. The market operates with a general price referred to as the "system price" which is (approximately) the price in unconstrained dispatch. Estimated congestions at the time of market clearing are mainly managed through zonal pricing, (see Bjørndal and Jörnsten (2001)). With the capacity charge approach, however, capacity charges are announced prior to market clearing, and for each node instead of only for a few zones. Moreover, the use of nodal capacity charges is similar to the pricing of marginal losses where loss factors are published for every connection point in the central high voltage grid. We also believe that the suggested procedure will easily facilitate bilateral trading to go alongside with the pool. As for the iterative process, we note that many congestion situations seem to last for a period of time (new zonal divisions within Norway are, for example, defined only for lasting new capacity limitations). In this case, market responses of initial capacity charge estimates might be used to obtain better estimates in subsequent periods, and real-time balancing of the system is handled by the regulating power market, a market that is already present.

Further development of the capacity charge approach, however, still leaves a number of questions to be investigated. First, we would like to do numerical tests on larger networks, and the exact procedure used for adjusting shadow price estimates is of special interest. In this, we can rely on optimization theory, for instance to find an adjustment scheme producing near-optimal dual variables within a few iterations. In this setting, we should also consider how close to optimality we need to be in order for the system to perform satisfactorily. In this case the procedure has to be combined with some other mechanism to obtain feasible flows.

Moreover, it would be interesting to perform simulations, where market data are slightly perturbed in each step. This would simulate how market prices and congestion develop in an adjustment process where we use market information from similar congestion situations to obtain initial guesses on the shadow prices. Since the suggested procedure also involves a mechanism like for instance curtailment or counter trading to obtain feasible flows, the specific design of the mechanism is of special interest, also taking into account how it affects the performance of, for example, separate real-time balancing power markets. Furthermore, gaming possibilities and more generally, regulatory issues should be examined.

Employing the "DC" approximation of the power flow equations, we do not focus on transmission losses, although a complete system for transmission pricing should address losses as well. We believe that marginal losses can be readily taken care of in our approach by issuing nodal loss factors similar to what is already done in the present Norwegian system, but possibly with a "hub" where the pool is virtually located, as a reference point when computing loss factors. Finally, one of the other simplifications of the "DC" model is that load factors are constants. The non-linear nature of the AC power flows implies that loss factors depend on loads, however our approach is still valid if the load factors used are marginal load factors, (see Stoft (1998)).

References

Bjørndal, M., & Jörnsten, K. (2001). Zonal pricing in a deregulated electricity market. *The Energy Journal, 22*, 51–73.

Bjørndal, M., Jörnsten, K., & Rud, L. (2000). A price adjustment process for managing congestion in the Norwegian scheduled power market. *Conference Proceedings, IAEE Norge Conference*, August 2000.

Chao, H. -P., & Peck, S. (1996). A market mechanism for electric power transmission. *Journal of Regulatory Economics, 10*, 25–59.

Dolan, A., & Aldous, J. (1993). *Networks and algorithms: an introductory approach*. Wiley.

Everett, H., III (1963). Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research, 11*, 399–417.

Glavitch, H., & Alvarado, F. (1997). Management of multiple congested conditions in unbundled operation of a power system. *IEEE Transactions on Power Systems, 13*, 374–380.

Hogan, W. W. (1992). Contract networks for electric power transmission. *Journal of Regulatory Economics, 4*, 211–242.

Minoux, M. (1986). *Mathematical programming. theory and algorithms*. Wiley.

Schweppe, F. C., Caramanis, M. C., Tabors, R. D., & Bohn, R. E. (1988). *Spot pricing of electricity*. Kluwer.

Stoft, S. (1998). Congestion pricing with fewer prices than zones. *The Electricity Journal*, (May), 23–31.

Wu, F., & Varaiya, P. (1995). *Coordinated multilateral trades for electric power networks: theory and implementation*. Department of Electrical Engineering and Computer Sciences, University of California.

Wu, F., Varaiya, P., Spiller, P., & Oren, S. (1996). Folk theorems on transmission access: proofs and counterexamples. *Journal of Regulatory Economics, 10*, 5–23.

Harmonizing the Nordic Regulation of Electricity Distribution

Per J. Agrell[1] **and Peter Bogetoft**

Abstract Regulators for electricity network infrastructure, such as electricity distribution system operations (DSOs) face some particular challenges in the Nordic countries. Due to institutional, economic, and historical reasons the DSOs in the Nordic area are relatively numerous and heterogeneous in terms of ownership structure, size, and operating conditions. Since the deregulation in 1994–1999, the national regulators have independently devised regulation mechanisms that address the heterogeneity through econometric or engineering cost models as a basis for high-powered regimes. The frontier analysis models (such as data envelopment analysis in, e.g., Norway and Finland) are particularly useful here, given their incentive properties and cautious estimation of the production set. However, the total information rents in yardstick regimes and the bias in the frontier estimation are related to the number of observations (firms), which undermine their future application in the Nordic area under increasing interregional concentration. This paper develops a proposal for an alternative model, the revenue yardstick model, that can be applied across the national regulations and permit frontier estimations on final user cost rather than cost estimates, sensitive to, e.g., capital cost estimates, periodization, and allocation keys. The core of the model is a dynamic frontier yardstick model such as Agrell et al. (2005), but here applied only to strictly exogenous conditions, the output dimensions and the claimed revenues of the DSO. An equilibrium is implemented using asymmetric penalties for positive and negative deviations from the ex post frontier revenue, the yardstick, using the classic super-

[1]The constructive comments from two anonymous referees on a previous version of the manuscript are acknowledged. Support from Nordenergi for the NEMESYS project and contributions by Sumicsid AB, EC Group AS, Gaia Group OY, SKM Energy Consulting AS and RR Institute of Applied Economics as project partners are acknowledged. The paper represents only the opinion of the authors that remain responsible for any errors in the presentation.

P.J. Agrell (✉)
Louvain School of Management and CORE, Université catholique de Louvain,
1348 Louvain-la-Neuve, Belgium
e-mail: per.agrell@uclouvain.be

P. Bogetoft
Department of Economics, Copenhagen Business School, 2000 Frederiksberg, Denmark
e-mail: pb.eco@cbs.dk

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*,
Energy Systems, DOI 10.1007/978-3-642-12067-1_17,
© Springer-Verlag Berlin Heidelberg 2010

efficiency model in analogy with Shleifer (1985). The model is particularly aimed at an international (interregional) application as it may embed national differences in regulation without jeopardizing the long-term sustainability of the model.

1 Scope and Introduction

1.1 Background

The objectives of this paper are to:

- Evaluate the advantages and disadvantages of a pan-Nordic regulation model and benchmarking tools viewed from all perspectives of the stakeholders, i.e., customers, regulator, owner, and distribution system operator
- Identify the most critical factors in cross-border regulation and benchmarking
- Present a common model for regulation of electricity distribution companies

The paper is interesting from several theoretical viewpoints. First, it proposes a partial harmonization of the national regulation mechanisms in anticipation of problems related to data access and consolidated operators. This problem has not been acknowledged in the scientific literature. Second, it extends the stream of literature on dynamic yardstick regimes, following the seminal papers by Shleifer (1985) and Meyer and Vickers (1997), among others. The proposed model uses an *ex post* correction of collected revenues rather than estimated costs, which has not earlier been studied. Although requiring a set of penalties to mitigate distortions and collusive outcomes, the regime offers advantages under international network regulation, when heterogeneous capital valuation techniques and accounting information convey less information than prospective investment information.

From a policy perspective, the proposal evokes the discussion on the challenges related to harmonizing network regulation for various stakeholders, even when empirical work shows strong consistency in client preferences. It also provides a rare example of regional industry initiative, the NEMESYS project as being research commissioned by Nordenergi, the industry association for electricity sector in the Nordic countries, for a common regulation model for electricity distribution in the Nordic region (NordPool region). Clearly, it is the improved incentive properties and the closer resemblance to competitive mechanisms that have persuaded the industry. On the regulatory side, the still open question is to which extent regulatory harmonization is compatible with the current institutional setting for the integrated energy market.

Although we have chosen to present the framework uniquely in the context of economic regulation of distribution system operators, there are also analogies and synergies in the specific Scandinavian segment of regional transmission operators. Some ideas on the potential harmonized regulation between (in particular) Norway and Sweden are developed in Agrell and Bogetoft (2008) and left outside of this paper.

2 Harmonization of Regulatory Systems

2.1 Current Situation

In NEMESYS (2005a), we demonstrated how the current regulatory systems for distribution system operations (DSOs) are based on somewhat different mechanisms. The regulation systems in each country can be summarized as follows:

- Denmark abandoned in 2003 the somewhat complicated revenue cap and rate of return regime first introduced in 2000, and moved to a temporary price fixation scheme with the aim to construct a new price cap system and settle issues regarding capital ownership.
- Finland has a well-established rate-of-return *ex post* approach. Since 2005, it has been refined and complemented with an *ex ante* cost cap component of CPI-X type. A company specific X-factor based on data envelopment analysis (DEA) and stochastic frontier analysis (SFA) models was introduced for the 2008–2011 regulatory period (EMV 2008).
- Norway has adopted a CPI-X type of revenue cap approach with clear *ex ante* emphasis. The system is established and stable. DEA benchmarking (yardstick) is used for defining the company specific X factors intended to be more important from 2007, NVE (2007).
- Sweden has moved from the light-handed *ex post* regulation (and an interim price freeze) to use of *ex post* technical norm model (network performance assessment model, NPAM) during the period. Concession granting is seen as a long-term component in the regulation. A DEA based benchmarking serves for information dissemination purposes. The regulation will undergo a major reform in 2012 (SOU 2007).

This illustrates that even though the systems aim at rather similar goals (creating markets in production and sales and guaranteeing reasonable tariffs), they are philosophically and technically somewhat different. Table 1 summarizes the different approaches used in regulation in the four countries.

The main philosophical difference is probably between Sweden relying on a light-handed *ex post* approach and Norway relying on a somewhat heavy-handed

Table 1 Summary of regulatory approaches used in the Nordic countries and year of deregulation

	Denmark	Finland	Norway	Sweden
Regime	1999	1995	1990	1996
Light-handed				1996–2002
Rate-of-return	2000–2003	1996–2004	1990–1996	
Cost cap		2005–2011	(2012–?)	
Price cap	2003–2007			2001–2002
Revenue cap			1997–2006	
Yardstick DEA			2007–2010	
Yardstick comp.				2003–2011

ex ante regulation. Still the difference between the *ex ante* and *ex post* perspectives are not always large in practical implementations – and EU regulation is calling for a common approach emphasizing *ex ante* applications. Technically, the main difference is probably between the Swedish network performance assessment model (NPAM) and the empirical frontier models used in the other countries. Again, however, the complementary DEA model in Sweden has many similarities with the benchmarking model in the other countries, even the Danish one that relies on a simple variant, the so-called COLS approach.

The countries are on very different stages in the implementation of the regulation systems. Both Sweden and Denmark have experienced problems with their regulation systems, and this has resulted in changes in the regulation principles. The Swedish regulation is about to change radically; the proposal SOU (2008) for 2012 advances abandoning the controversial NPAM in favor of rate-of-return regulation on capital and a cost-yardstick regime for controllable operating expenses. Norway has proceeded relatively consistently with the same approach. The Finnish situation is somewhere between the extreme cases. In spite of open and frequent information exchange between the Nordic regulators in FNER/NordREG and bilaterally, there has been no natural harmonization of the systems.

Although there have been some attempts to coordinate certain tasks, such as the Nordic DSO benchmarking 2001 (Førsund and Edvardsen 2001), data sharing is of limited usefulness as long as the DSO tasks are somewhat different without any estimate of their relative or absolute importance. There are also more practical differences even on the information collection level. Due to historical reasons, division between transmission, regional networks and distribution differ (voltage levels) varies. There are also many other smaller differences in the ways the key indicators are defined. As the time lags in the collection of data are long, this is one practical issue that hinders harmonization and even less formal benchmarking, etc.

2.2 Effects of Harmonization

The *advantages* of harmonization of regulation are effects inducing:

- Improved long-term stability and hereby protection of specific investments by making the commitment at an international level
- Better structural adaptation by making it easier for DSOs to operate in different countries, by compensating for the small sample bias problem, by avoiding that the regulator reduces mergers, etc., to keep a sufficiently large number of observations
- Improved learning across DSOs, regulators, and other stakeholders
- Increased competition for DSO role by making it easier for DSO to offer services in different countries
- Increased EU influence by taking the EU initiative in terms of network regulation, energy market design, and coordination

- Improved competition at the suppliers' level – e.g., among equipment manufacturers, constructors, and software providers that have to only learn one rather than multiple regulations, i.e., low barriers to entry and greater returns to scale on specific investments upstream
- Less trial and error by learning from best practice in regulation and from pooling regulatory resources
- Improved foresight and predictability for network users, e.g., with respect to tariff development, investments, and quality
- Lower administrative costs for the regulation in the reporting, accounting, etc
- Lower costs for regulators to refine tools and instruments for a common model
- Symmetric Nordic coordination of the electricity sector via a semistructured coordination at DSO level, matching well the existing coordination at TSO and generation levels

There are also some *limitations* to the harmonization approach:

- Stranded investments adapted to regulation that is discontinued or radically changed, e.g., specific assets or processes
- Risk of hampering regulatory innovation
- Time spent to set up and launch legal and institutional obstacles
- Contradictions with current Electricity Acts, preambles, or other laws
- Internal DSO conflicts may overlook regional differences

Hence, we should strive for an optimal level of harmonization: touching the principles of regulation as to gain commitment; defining a set of common tasks as to improve data quality; and gathering cost information, creating sound and equal incentives of investment, and efficient operation across the region. The harmonization should not limit national prerogatives unnecessarily, force potentially arbitrary institutional reforms upon countries or sacrifice rational national adaptations to operating conditions in order to achieve common standards.

2.3 Analyzing the Nordic Case

The Nordic area shows, as argued above, signs of suboptimal costs of regulation at several levels. First, the tasks are defined nationally, if at all, leading to high coordination and information costs. Second, the due to the different approaches and stages of implementation, the Nordic regulation systems provide quite different incentives for DSOs. The incentives for efficiency improvements depend heavily on the possibility to gain additional profits by making improvements that exceed expected levels. In Finland and Norway, this is possible during the regulation period, but the effect on the base line for the next period gives a mixed signal (the *ratchet effect*). The effect on the tariff level depends on the tightness of regulation, the clarity of the requirements *ex ante*, and the obligation for the return of excess profits to the customers. For example, in Finland the last two aspects have been changed and the incentive for tariff changes has increased significantly. None of the countries offer very clear

incentives related to security of supply or other quality issues. This reflects the fact that historically the quality performance has been perceived as good. None of the current models provide any clear (wanted on unwanted) signals for consolidation. Finally, and most alarming, the current changing regulatory landscape does nothing to address the poor investment incentives in the sector, where the regulatory risk is evaluated strictly nationally even for consolidated firms.

2.4 NEMESYS View on Harmonization

A process of harmonization could significantly improve the network regulation based on the analysis above for the Nordic scenario. A harmonization would require two necessary stages: a definition of the scope of the common and individual tasks and the elements of the common regulatory framework. Before outlining the stages, let us quickly clarify some aspects that are not in the scope of this paper:

- No creation of regional or European meta-regulator. We acknowledge that network regulation outside of the studied area may be heterogeneous for other reasons than those investigated and, furthermore, the institutional complexities involved in the centralization of legal powers contrary to the subsidiarity principle of the Moreno doctrine (cf. Majone 1997, for a discussion of this issue).
- No imposition of common tariffs or delivery conditions across regions. In choosing the term *harmonization* rather than *standardization*, we acknowledge that the principal benefits from a common regulatory framework do not derive from identical tariffs (in fact, the Nordic area has always had individual tariffs even within jurisdictions), but from a consistent set of regulatory expectations and contracts.

2.4.1 Definition of DSO Task Portfolio

Any coordinated regulation system requires a clear definition of the scope of the tasks to be performed by the regulated firms. In the case of a harmonization, i.e., the creation of a common set of regulations and a set of compatible regulations for other tasks, this definition is primordial. Let us call the common definition the *Core DSO Task* (cf. NEMESYS 2005c). Although the definition may be done as the minimal intersection of all national DSO tasks, the true implication is that information acquisition should be decomposed in the core task to allow for straight-forward comparisons, which might require some changes of the national reporting systems.

– *Core DSO Tasks* form the least common subset of DSO tasks to be fulfilled by any DSO in the Nordic area and for which common data is collected and common regulation could apply.
– *Nationally Regulated DSO tasks* are defined DSO activities that are not harmonized in the Core DSO Task but that remain regulatory obligation in at least one country. This could apply to, e.g., tasks related to safety inspections, line dismantling and energy planning. All compensation for such tasks should be transparent

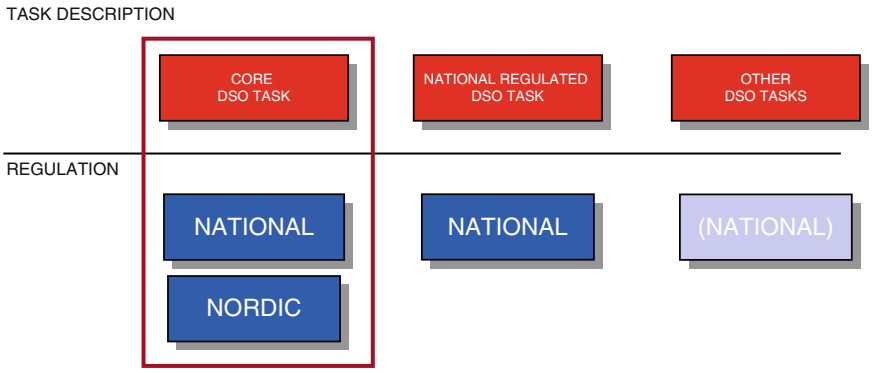


Fig. 1 DSO tasks and corresponding regulation

(and preferably based on tendering) to avoid cross-subsidies to incumbents. To maintain this transparency and promote development, we propose that the regulated payments for the National Regulated DSO Task should be separated from the Core DSO Task. In practice, this could render these tasks more attractive to non-DSO providers or at least informing the regulator of the real costs involved to permit social trade-offs.

– *Other DSO Tasks* are activities that are not regulated, but compatible with the national and European directive with respect to nondiscrimination, independence and competition. In the regulatory regime proposed below, such competitive activities can be freely performed under residual competition law so far as they bring coordination gains (Fig. 1).

2.4.2 Regulation and Information Harmonization

– *Framework agreement*, a common regulatory vision statement for both model structure and the time plan need to be agreed upon by all regulators. This does not mean a streamlined legislation, but a high-level commitment to the principles and tools for all further revisions. A fully integrated regulatory body like NORDREG could easily administer the regulation, but the proposal is flexible with respect to institutional solution. The competency to define national parameters (e.g., interruption costs), nationally regulated DSO tasks and to define concession areas is also a national prerogative, as all monitoring of the noneconomic and equity aspects of the directive.

– *Common information system*. Any quality-oriented, output-based regulation needs access to high-quality data in common formats, but so does the sector itself for its coordination and restructuring. We propose a common client metering standard, including format for transmission of hourly data, connection and disconnection. We propose that the meter standard defined by a *Metering Agent*, formally responsible for the metering, reporting, and administration of technical data from customer level to other parties.

– *Transparency in financial conditions*. To increase transparency and avoid regulatory competition, the regulators need to coordinate the financial conditions also for Nationally Regulated DSO Tasks. However, this should be seen in connection with a common information system, leading to equal performance criteria in, e.g., reliability and commercial quality. Note that this does not mean that the average realized profits need to be equal in all countries, since they are the outcomes of the regulation itself.

3 Regulatory Toolbox

This section draws on a richer discussion on design of economic regulatory instruments for network regulation in NEMESYS (2005b). The practice and theory of network regulation exhibits a plethora of models and approaches, which basically can be classified into five categories of regimes:

– *Cost-recovery*. A rate-of-return, cost-oriented regime such as widely present in USA and earlier in, e.g., Finland and Norway. Firms are authorized to a predetermined capital cost on pre-approved investments in addition to direct by-pass of certain operating costs.
– *Price cap*. A regime of the Anglo-Saxon CPI-X type (cf. Littlechild 1983) where the regulator *ex ante* determines a fixed reduction (X) of some base-level price or revenue. In practice, the regulatory asset base is approved by the regulator that also decides on the return for investments in the period. At the end of some period, the base is reset to current cost.
– *Yardstick regime*. A regime in which costs or revenues are set competitively across comparable firms by using averages, best practice or frontier models. Allowable revenues next period depend on performance the previous period. For electricity distribution, the models of the Agrell and Bogetoft (2005) type are considered due to their capacity to cope with multi-dimensional performance under weak prior technology assumptions.
– *Franchise auctions*. Firms are awarded concessions based on tender auctions formulated in tariff level using a pre-defined task specification. The concessions are defined on time periods between 5 and 15 years and then resubmitted. The proposed model in NEMESYS (2005b) draws on the contestable licensing model in Neeman and Orosel (1999), but the seminal reference for infrastructure franchises is Williamson (1976).
– *Technical norms*. The allowable revenues for the firms are determined using an engineering cost model. The model presumes a complete task description, including quality provision and technical development.

NEMESYS (2005b) evaluates the five categories with respect to the overall criteria to find guidance in Table 2 for the development of interesting candidates for a common regulation regime. Green areas indicate relative strengths, red indicate weaknesses, and yellow the aspects dependent on parameters. Two findings are

Table 2 Comparison of model alternatives

Concern	Cost-recovery	Price/revenue caps	Yardstick regimes	Franchise auctions	Technical norms
Optimal allocation of decisions and information	−	−	+	+	−
Incentives for sound industry structural changes	−	+	+	+	−
Incentives for efficiency improvements	−	0	+	+	+
Incentives for tariff reductions	−	0	+	++	−
Incentives for service quality improvements	+	−	?	−	−
Incentives for (re)investments	+	?	?	++[2]	−
Long-term regulatory credibility	0	0	+	0	−
Unbiased DSO performance assessment	+	+	+	?	?
Low administrative costs of regulation	−	0	0	−	−

[2]The investment incentives under franchise auctions depend on the number of firm bidders (supply market), the information available for and among operators about the operating conditions and costs in the concessions (information asymmetry) and the commitment and credibility of the regulator to keep the settlement (regulatory uncertainty). Here we assume that the bargaining power will be stronger at the firm level and that the regulators would not renege on their contracts.

apparent: the behavioral advantages of yardstick and franchising systems and the specific need to address service quality in the regime. On the one hand, the two methods bridge the information gap between the regulator and the firm, in that they form a "pseudo"-market for the firms. This allows the regulator to concentrate its efforts to areas where it is necessary and relevant, such as monitoring of terms, industry structure, and quality development. On the other hand being highly incentivized, their effectiveness depends crucially on the regulatory commitment, which is where the regulatory integration comes into play. A specific section is devoted to the economic regulation of service quality as a consequence of its *public good* properties and the interest to create explicit incentives for continued investments in service provision. The investment incentives crucially depend on the parameters and regulatory commitment to the system, they can be either very good or very poor. In NEMESYS (2005b), this analysis is taken as a starting point to develop in more detail alternatives for quality, yardstick, and franchising regimes.

3.1 Dynamic Cost-Based Yardstick

Dynamic yardstick schemes based on DEA solve many of the usual CPI-X problems, including the risk of bankruptcy with a too high X, risk of excessive rents with a too low X, ratchet effect when updating X, arbitrariness of the CPI measure, arbitrariness of the X parameter, and inability to include changing output profiles. The most important difference between a yardstick schemes and a more traditional CPI-X regime is that the firms are compared to actual cost frontiers rather than projected cost frontiers. The incentive for an efficient firm to improve performance is assured by using the so-called *superefficiency* specification (e.g., Bogetoft 1997), whereby the frontier for each firm excludes its own observation. This reduces the informational and analytical requirement put on the regulator and allows for a more precise inference of actual performance. It hereby also allows for better incentives. The business risk is not increased as revenues follow more closely the cost development.

3.2 Dynamic Revenue-Based Yardstick

To relax the input-dependency of the previous regime and in particular to alleviate the capital evaluation problem an innovative model based on net revenues rather than costs is developed in NEMESYS (2005b). What matters to consumers is the *final tariff* = *revenue*, not the *cost* as such. Using actual unregulated prices in the yardstick the return on investment is endogenous and not regulated. DSO charges are set by all firms and regulated afterwards depending on the "value for money" set by the other firms. In this way, firms may budget for reinvestments prior to investment, rather than getting caught up in the jerky and artificial problem of network age.

The basis for the dynamic revenue-based yardstick is still a benchmarking or frontier model that calculates for a given level of output in all its dimensions, taking into account the operating conditions, efficient revenue by a comparable operator. The positive or negative difference between the efficient revenue and the actual revenue charged becomes a *carry-forward* that is to be repaid or charged with interest rates, just like a loan to and from the rate-payers. Since the total relevant cost includes all operating, capital and financing charges, cost pass-through can be limited to standard costs for network losses, transmission charges, nondistribution tasks, and taxes.

3.3 Dynamic Network Auction Model

The strength of the yardstick model is also its weakness in the long run when the market consolidates. The annual revisions become ineffective if only very few firms compete on the market. In fact, the uncertainty shifts to the clients, risking to pay all cost shocks and the rents of market power. To address this situation, one may augment the yardstick model with a fixed element, which also solves possible

imperfections in the environmental correction. A new model in NEMESYS (2005b) solves the residual benchmarking estimation problem by a repeated auction design.

Briefly, in the dynamic network auction model the DSOs tender for the fixed tariff in their concession area, knowing that their variable tariff will be regulated as in the dynamic revenue yardstick above during the duration of the license. In this way, the consumers are assured to get higher competition for the market, while still being protected from any attempt to recover an artificially low fixed tariff by an inflated variable tariff. On the other hand, the firm is protected by the fixed component against possible systematic errors (up or down) in the yardstick model. In case the operating conditions are understated in an area, the DSOs would require an extra payment; if they are less severe they may even pay to get the license.

3.4 Quality Regulation Model

As discussed above and in NEMESYS (2005a,b), the quality dimension is ever more important for the network regulation at all levels. The ability of the regulation to adequately and credibly provide incentives for long-run quality provision will be one of the acid tests for the regulation. First, supported by the theory and the scientific consensus of SESSA (2005), we conclude that a regulation for electricity distribution that is entirely restriction-based is likely infeasible in the long run. However, the large number of measurable dimensions suggests a hybrid approach using restrictions, since many of them are correlated to reliability of supply. Thus, we argue for the explicit inclusion and marginal pricing of reliability of supply, such as is currently done in Norway, Germany, and Sweden. Other quality aspects, related to voltage and commercial quality, may be defined with target and threshold values in the DSO Task Description, preferably jointly with clients and industry organizations.

4 The NEMESYS Approach

The approach, fully documented in NEMESYS (2005d) is composed of two elements: the *Revenue Yardstick Model* and the *Quality Incentive Scheme*.

4.1 Revenue Yardstick Model

The yardstick model is founded on the virtues of yardstick competition, i.e., the DSOs can compete even though they do not meet directly at the market. This safeguards the consumers against too high tariffs and it safeguards the DSOs against unreasonable impact from regulatory interference based on limited information. The economic condition of one DSO is basically defined by the other DSOs, not by a regulator.

The revenue yardstick model defines the *revenue base* $RB(t)$ for a given DSO in period t as

$$RB(t) = C^*(t-2)$$

where $C^*(t-2)$ is the yardstick revenue for period $t-2$ determined by the benchmark model estimated on the data from all other DSOs except the one in question (superefficiency evaluation), cf., below. Invoking a 2-year delay enables (1) the DSOs to do their financial accounting in the usual way, (2) the regulator to collect and process tariff and service data in time, and (3) the consumers to know tariffs a priori.

The (benchmarked) *DSO charges* in period $t-2$, $C(t-2)$ may deviate from the yardstick revenue. If these DSO charges have exceeded the yardstick revenue, it corresponds to the DSO having taken a loan with the consumers. If it falls short of the yardstick revenue, it corresponds to the DSO having provided a loan to the consumers. These loans should be repaid with interest.

We shall think of these as *carry-forwards* in period t, $CF(t)$, i.e., we have

$$CF(t) = \begin{cases} (1+\alpha)\left[C^*(t-2) - C(t-2)\right] \text{ if } C^*(t-2) \geq C(t-2) \\ (1+\beta)\left[C^*(t-2) - C(t-2)\right] \text{ if } C^*(t-2) < C(t-2) \end{cases}$$

The parameter α is the two-period borrowing interest rate in period $t-2$ applying to the first case of *undercharging*. The second case in *CF* is the *overcharging* which is penalized with a is a lending rate $\beta = \alpha + \delta$ that exceeds the two period costs of borrowing with some extra margin $\delta > \mathbf{0}$. In the following, we shall think of a period as 1 year.

The sum of the revenue base and the carry-forward defines the *revenue target* for period t as

$$RT(t) = C^*(t-2) + CF(t).$$

The revenue target *RT* is indicative as a budget for tariffsetting. It defines the actual charges the DSO in question should make in period t to come out on equal footing with the other DSOs presuming that they do not change from period $t-2$ to period t. The indicative revenue target can be used by the regulator when ruling on or confirming actual charging proposals $AC(t)$ for period t at the end of period $t-1$, cf. below. Exactly how the regulator rules here is not very important for the incentive properties of the scheme and the regulators in the different countries need not even use the same principles. What is important for the convergence and the compatibility with the Directive is that the methodology for determining the revenue yardstick and target is defined *ex ante*.

In period t the actual charges of the DSO is denoted $AC(t)$.

The actual charges will, however, reflect not only the costs and profits to the DSO in period t but also the settlements of negative or positive carry-forwards. Therefore, the real in-period DSO charges in period t, the benchmarked charges $C(t)$ is

$$C(t) = AC(t) - CF(t)$$

The benchmarked charges form, together with the provided services, the basis for the benchmarking exercise that set the revenue base $RT(t+2)$ for period $t+2$, i.e., $C^*(t)$.

4.1.1 Example

To illustrate the mechanics, consider a case with three DSOs that have so far in each and every period charged their full cost ($C(1) = 100$). Let the interest rate be $\alpha = 5\%$ and the penalty rate $\delta = 5\%$. The development in underlying minimal costs is illustrated in *italics* and the actual charges are given as AC in Table 3 below. DSO One and Three realize a cost decreasing change to lower their cost to 90 in Period 2. We see that DSO Two faces idiosyncratic extra costs of 10 in Period 2 and that DSO Three uses its relatively low costs to extract extraordinary profits in Period 2. The decision not to lower the tariffs to the optimal level by DSO Three is then providing the incentive for DSO One, a positive carry-forward that is collected in Period 4 (with interest).

The example not only illustrates the formulae above. It also illustrates that companies carry their idiosyncratic risks, are ensured against general variations in costs and that there is pressure on the DSOs to reduce charges to the minimal level

Table 3 Example of revenue yardstick regime, five periods, three firms

		Period 1	Period 2	Period 3	Period 4	Period 5
DSO One						
Yardstick revenue	$RB(t) = C^*(t-2)$	100	100	100	100	90
Carry-forward	$CF(t)$	0	0	0	10.5	0
Total costs	$c(t)$	*100*	*90*	*90*	*90*	*90*
Actual charges	$AC(t)$	**100**	**90**	**90**	**100.5**	**90**
Benchmarked charges	$BC(t) = C(t) = AC(t) - CF(t)$	100	90	90	90	90
Extraordinary profit	$AC(t) - c(t)$	0	0	0	10.5	0
DSO Two						
Yardstick revenue	$C^*(t-2)$	100	100	100	90	90
Carry-forward	$CF(t)$	0	0	0	−11.5	0
Total costs	$c(t)$	*100*	*100*	*90*	*90*	*90*
Actual charges	$AC(t)$	**100**	**100**	**90**	**78.5**	**90**
Benchmarked charges	$BC(t) = C(t) = AC(t) - CF(t)$	100	100	90	90	90
Extraordinary profit	$AC(t) - c(t)$	0	0	0	−11.5	0
DSO Three						
Yardstick revenue	$C^*(t-2)$	100	100	100	90	90
Carry-forward	$CF(t)$	0	0	0	−11.5	0
Total costs	$c(t)$	*100*	*90*	*90*	*90*	*90*
Actual charges	$AC(t)$	**100**	**100**	**90**	**78.5**	**90**
Benchmarked charges	$BC(t) = C(t) = AC(t) - CF(t)$ $AC(t) - c(t)$	100	100	90	90	90
Extraordinary profit	$AC(t) - c(t)$	0	10	0	−11.5	0

that covers all costs, including capital costs. Note the superefficiency method in operation, without which DSO One would have had no incentive to reveal the true cost information in Period 2.

4.1.2 The Intuition

The intuition of the revenue yardstick model is as follows: In period $t-2$, the DSO is first and foremost allowed the efficient tariff charges, $C^*(t-2)$. For practical purposes, however, the allowed income is determined based on data with a two period delay. This will allow the regulator time to collect data from period $t-2$ during the first half of year $t-1$, and to calculate the allowed revenue for period t during the last part of year $t-1$. The DSO and regulator can therefore settle on period t charges a priori. This has two advantages compared to a direct implementation of the revenue yardstick without time-delay. First, it allows a DSO to close its financial statement according to normal procedures. Secondly, it ensures that the regulation complies with even strict interpretations of the *ex ante* provision in the EC directive.

The model works with asymmetric interest rents. Undercharging carries the normal interest rate α. Overcharging must be paid back using a higher rate β. In principle, the scheme is incentive compatible even when lending and borrowing carry the same interest rate, $\alpha = \beta$, but to make the scheme more high-powered, and clear we propose to add an extra charge $\delta > 0$ in the case of overcharging. Coupled with the uncertainty of the yardstick level, this will give the DSOs extra incentives to reduce charges.

The revenue yardstick scheme is illustrated in Fig. 2 below. The minimal costs of an efficient DSO is indicated with a non-filled point below the yardstick level,

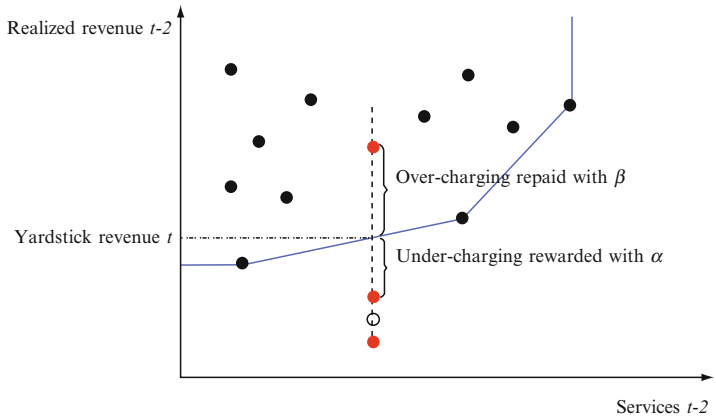


Fig. 2 Yardstick revenue scheme

but the DSO can choose to charge the consumers more or less as indicated by the solid points on the vertical line through the minimal cost point. Overcharging occurs when the charging exceeded the yardstick level.

4.1.3 Incentive Effects

The proposed scheme gives the DSOs incentives to participate and to reduce tariffs to the smallest level that is consistent with continued operation. In short, the mechanics is based on rational forward-looking decision making, where a DSO would not charge lower tariffs than what corresponds to continued operation, including reserves for future investments and a fair long-run average rate of return. The difference is particularly visible for old networks before reinvestment or expansion, where an input-based regulation (using some book-value estimate of capital) would artificially "strangle" the build-up of reserves prior to investment, later to hike tariffs up after the investment. In the new model, this problem disappears as it is the value of continued operation, the tariff, that is evaluated, not the age or state of its assets.

4.2 Benchmarking Model

The benchmarking model is the engine in the yardstick model to determine the revenue target C^* for any kind of DSO; rural or urban, with any kind of service profile, mix of high voltage/low voltage. NEMESYS (2005c) is entirely devoted to this model, for which we briefly touch the specification and the choice of model below.

In the *specification*, the model should take into account inputs, outputs and environmental conditions. On the input or cost side, we need the *revenue levels*. Since the total relevant cost includes all operating, capital and financing charges, cost pass-through can be limited to standard costs for net losses, transmission charges, nondistribution tasks, and taxes. On the output or *services* side the scientific as well as the technical literature converges on a specification that reflects three dimensions: customer service, transportation work, and capacity provision. The first dimension is usually measured by the total number of clients, potentially divided into voltage levels or market segments. The second corresponds to total delivered energy, if needed differentiated by voltage level. The third dimension is measured by proxies for capacity such as installed transformer power or peak power. Environmental conditions can be covered by network length (all studies), delivery area (UK), climate zone (previously in Sweden), or other proxies, cf. Agrell and Bogetoft (2003, 2005, 2007).

Concerning the *type of model*, we draw on the economic optimality and international experience of the DEA model for network regulation, already in regulatory use in Austria, Germany, Finland, Norway, and Sweden The model has the advantage of giving a conservative estimate of efficiency and draws on a solid production

economic base. However, other models can also be applied, from simple partial averages (€/kWh delivered, etc.), linear cost functions (e.g., based on simple linear regression), or more advanced frontier functions such as SFA (cf. Agrell and Bogetoft 2005, 2007), in use in Germany, Finland, and Portugal.

4.2.1 Quality Incentive Scheme

As discussed above and in Chap. 4, the quality dimension is ever more important for the network regulation at all levels. The ability of the regulation to adequately and credibly provide incentives for long-run quality provision will be one of the acid tests for the regulation. Three dimensions emerged from the analysis in Sects. 2 and 4 above: (1) the quality steering, (2) the information requirement, and (3) the timing of information and settlement (*ex ante*, *ex post*).

For reasons of visibility and commitment, we propose a strict application of an *ex ante* marginal pricing scheme (cf., Sect. 4) on reliability. That is, while the tariff levels should be regulated by a yardstick scheme with the advantages of *ex post* evaluations, we propose that quality is regulated using a strict *ex ante* approach. The *Quality Incentive Scheme* is based on data collected per customer segments for each operator on ENS and SAIFI, defined as

- *ENS* (Energy Not Supplied, GWh), defined at client connection level (<1 kV) for interruptions longer than 1 min, divided into notified and non-notified interruptions.
- *SAIFI* (System Average Interruption Frequency Index), defined as the number of sustained interruptions reported at distribution delivery point (<1 kV), irrespective of interruption time, divided into notified and non-notified interruptions.

The proposed scheme has some similarity with the Norwegian cost of energy not supplied (CENS) as originally described in ECON (2000). The Norwegian CENS regulation is updated from 2009, cf. NVE (2007). Compared to the Norwegian CENS regime, the quality regulation involves a series of important improvements to get a better adjustment of realized quality levels to the socially optimal ones. Similarly, the proposed regulation resembles that implemented in The Netherlands. The proposed approach also has similarities to the Swedish customer reimbursement system although care should be taken to avoid unnecessary administrative burdens.

4.2.2 Compensation Scheme

The structure of the quality incentive scheme is simple

$$Q = A + pq$$

where Q is the quality payment to the DSO, A is a fixed payment q is the supplied level of quality dimension and p is a vector of prices per output dimension.

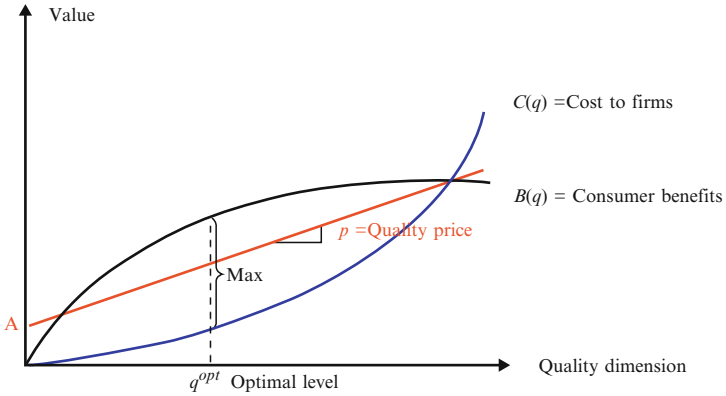


Fig. 3 Quality regulation

The quality reimbursement shall be added to the allowed revenue according to the revenue frontier yardstick model to form the total regulated revenue to the DSO. Specifically, this can be done by including the positive or negative quality charges in the carry-forwards into a carry-forward with quality CF_{wQ} so as to settle these with a 2-year delay.

– *Base payment A* is the preferred equilibrium point (base level) of ENS and SAIFI is to be determined once for each operator and concession area, without updating, using statistical, technical and socio-economic analysis. It shall reflect the underlying costs of quality provision and the socially preferred equilibrium values of ENS and SAIFI. In the analysis care should be taken to include the environmental, load and service factors that have an impact on historical reliability. The optimal levels, the base levels, trade off the benefits to consumers against the costs to the DSO of providing quality. The optimal levels will depend on the country, connection level and the DSO, but have no link to historical levels of cost or performance.
– *Marginal price p* is defined without ambiguity based on type of connection, for each customer segment by the regulators for a 10-year period (cf. Fig. 3) per unit of energy not delivered (ENS) and outage occasion (SAIFI), for notified and non-notified interruptions, respectively.

4.2.3 Regulatory Settlements

Based on objective and verifiable measurements of ENS and SAIFI at customer level compensations to individual customers can be calculated and the customers can be reimbursed with a time delay corresponding to the one used in the revenue regulation. To the extent that ENS and SAIFI cannot be measured and controlled at customer level, then the lowest, most customer close measure points shall be used.

In these points, the ENS price will equal an average of the consumer-based prices below while the SAIFI price will be the sum of the SAIFI prices for the customers below. To avoid unnecessary administrative burdens, small consumer reimbursements could also be accumulated on a solidarity account and be used to lower the general charges to the DSO's consumers.

4.2.4 Regulatory Procedure

The proposed regulatory procedure is illustrated in Fig. 4 below for the three parties *DSO*, *regulator* and *metering agent*.

1. *DSO* reports electronically financial, service and concession data for *Core DSO Tasks*, *National Regulated DSO Tasks*, and *Other DSO Tasks*, after closing the accounts for year 1.
2. *Metering Agent* reports low-level reliability data for year 1 and total supplied energy from higher grids.
3. *Regulator* calculates DSO Net Revenue by deducting for the preceding year, charges to higher grids, standard costs for network losses, *National Regulated* and *Other DSO Tasks* from the submitted total revenues. If the Net Revenue is negative, the firm is deleted from the list of comparators, otherwise not.
4. *Regulator* runs revenue yardstick model is run for all firms, using the eligible comparators, and the Efficient Revenue for year 1 is calculated for each firm.

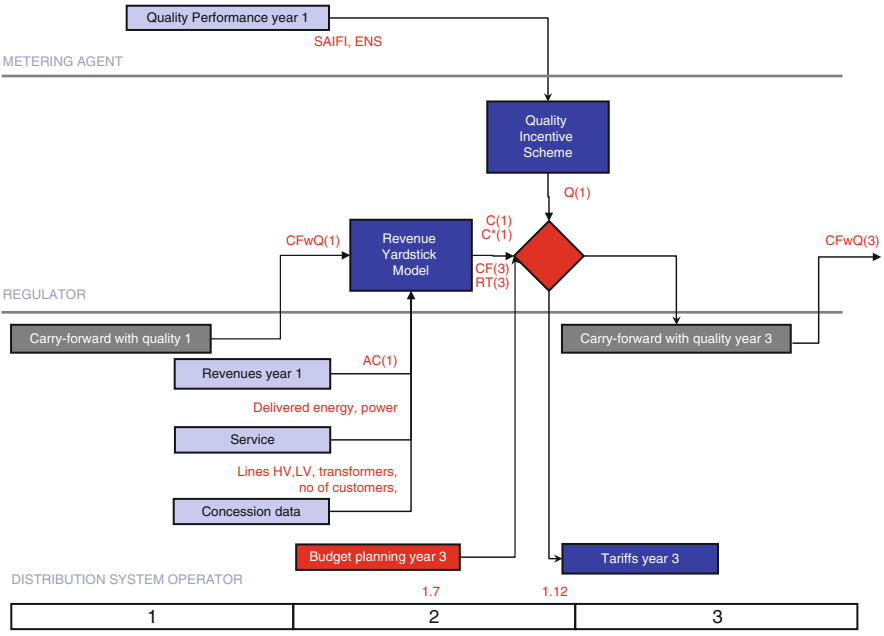


Fig. 4 The NEMESYS regulatory process

5. *Regulator* calculates the quality incentive scheme and the result can be either negative or positive. In the first case where less than optimal quality has been supplied, the result shall be processed as an ENS compensation to be paid out to consumers in year 3. To limit the administrative burdens, we may choose only to compensate individuals when the compensations exceed a given threshold, and to simply pool small amounts and use these to compensate the collective of consumers. In case of positive quality outcomes, i.e., when the supplied quality level exceeds the base level, we propose to simply charge the consumer collective in year 3.
6. *Regulator* announces the carry-forward for year 1, including the negative or positive difference from the revenue yardstick and the pooled difference from the quality incentive scheme, adding *National Regulated Tasks* as the settlement for the DSOs in year 3.
7. *DSO* incorporates the carry-forward for year 1 in the establishment of tariffs and a projected budget for year 3. The tariffs and the projected budget are submitted to the regulator as an acknowledgment of the carry-forward.
8. *Regulator* formally approves (orthodox *ex ante*) the proposed tariffs for year 3. In this step, discretion may be exercised based on, e.g., upcoming investments in the budget or negative carry-forwards.

5 Stakeholder Analysis

To investigate whether the institutional differences in regulation reflects underlying national differences in preferences and prioritization, the project collected structured data on the goals and objectives of different stakeholders (DSOs regulators, government, industrial clients, residential clients, associations) in the four countries. The study NEMESYS (2005a) comprised both questionnaires and semidirected interviews of the group.

5.1 Results: Stability, Quality, and Efficiency

The survey and interview material show that different stakeholder groups are to a large extent unanimous about the goals and objectives related to electricity distribution (cf. Fig. 5). Most important economic aspects are stable tariffs (clients) and stable return on investment (owners), and high efficiency. Natural conflicts are related to the level of tariffs and profits. Quality aspects are as important as the economic aspects and security of supply is the most important single goal. Other studies TNS Gallup (2005a,b) confirm increasing orientation towards customer service and reliability of information from the sector. Many clients are also uneasy with the basis and justification for the distribution tariffs, suspecting inefficiency. Equity and fairness issues, in particular open network access, are mostly seen as requirements

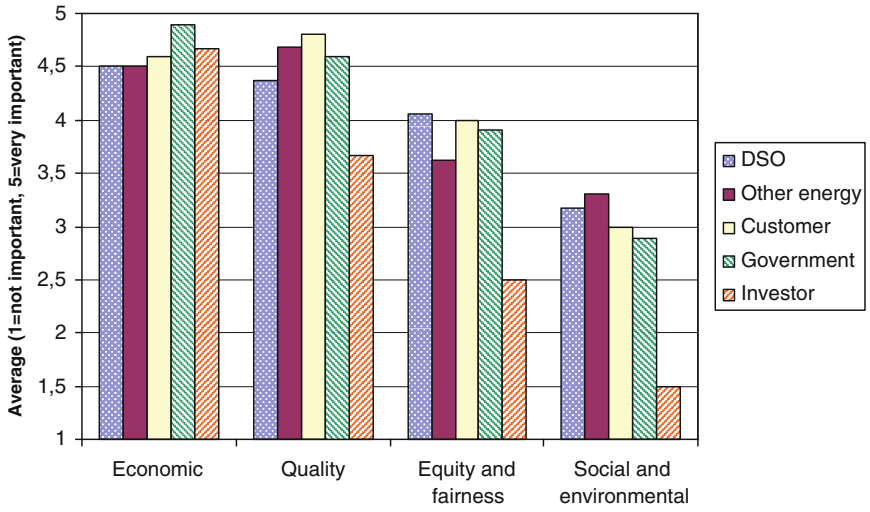


Fig. 5 Relative importance of DSO service dimensions under well functioning regulation (NEMESYS 2005d)

but not primary goals. Depending on the stakeholder groups, equity of customers and fairness for different types of companies are also seen important. Social and environmental aspects are ranked less important for network regulation.

5.1.1 Stakeholder Consequences

Summarizing a richer consequence analysis for firms, customers, market, and government in NEMESYS (2005d), the most important changes would be related to the incentives for lowering tariffs and improving cost efficiency. The suggested mechanism would set a high cost reduction and tariff cut pressure on those companies that are classified as inefficient. This would have a very significant local effect, and it would probably lead to structural changes in the industry. On the other hand the efficient companies would probably make more profit than under the current regimes. Hence, the suggested approach would give stronger incentives for improving efficiency and assuring capital for investment in grid expansion and reinforcement. In the short run, the suggested mechanism would decrease the stability of tariffs and profits, but in the long run it should lead to a stable situation.

5.2 Further Work

Many detailed questions remain before an actual implementation of either harmonized regulation or the specific proposal. Below we specify a prioritized list of suggested further work. Note, however, that these projects in no way can or should

substitute for development work made directly by, or for, the regulators. The active involvement of the sector in the development and convergence of regulation enables and facilitates the necessary implication of the national institutions. To further analyze the properties of the proposal we propose work on three themes: internal evaluation, enablers, and principal challenges.

5.2.1 Consequence Analysis

A harmonization like the one proposed in the paper is also associated with direct costs related to the change of system at a regional scale. It is, thus, indispensable to anchor the reform among the stakeholders and to evaluate the consequences in order to define the optimal scope and scale of the harmonization.

- *In-dept quantifications of the consequences* for the grid companies of changing the regulatory design from the current national regulatory models to the by the working group preferred Nordic regulatory model. The analysis should include quantifications for different types of grid companies as well as an industry-wide quantification.
- *Live simulations* could be done to explore the yardstick logic using a "virtual laboratory" with real decision makers subject to a dynamic simulation. The purpose of these experiments is to validate whether the decision makers perform according to the assumptions of the *new logic* and to collect data on behavior to facilitate parameterization.

5.2.2 Initiative on Regulatory Enablers

From the harmonization enablers mentioned above, further work is needed in:

- *Task definitions* among the membership countries as to determine a possible set of core DSO Tasks, including a set of indicators to measure the output of the tasks if necessary.
- *Data harmonization* should be defined in collaboration with regulators as to converge on data collection routines, metering standards, information systems, definitions, and standards that permit pan-Nordic open information systems.

5.2.3 Principal Challenges in the Model: Non-Profit Firms

The application of a high-powered regime (such as the proposed yardstick) to non-profit maximizing firms would need specific instruments to control for their impact on the revenue norm. A specific study should be conducted on (1) the prevalence and behavior of nonprofit firms and (2) an in-depth investigation to validate the accommodation of such firms in yardstick competition schemes.

6 Conclusions

The NEMESYS proposal is an innovative attempt to design a regulatory approach that improves on the most important dimensions for the Nordic stakeholders, the incentives for investment and efficiency, stable tariffs, and quality of service. The proposed approach differs from existing regulation in detail, but primarily in philosophy, as it is a consistently output-based regulation that delegates the process to the regulated firms. In doing so, it changes the information requirements in the regulatory approach in the direction of increased attention to what really matters to the final consumer, i.e., a clear and consistent description of the regulated task and performance assessment. It also constitutes a true paradigm shift in that it restores the role of the regulator to market design and surveillance of structure and development, rather than direct negotiation partners in a proxy-bargaining process on behalf of the customers. Hence, the competition in the NEMESYS approach is played between *firms* in operation achieving stable and low tariffs at high-quality, not towards the regulator using asymmetric information on current and upcoming investments.

The *harmonization* of DSO tasks, defining responsibilities, data and compensation norms clearly across the Nordic countries leads to a more transparent and constructive dialog with regulators. Further harmonization of regulation principles, instruments and regimes improves both the *effectiveness* of the regulation, by commitment to the key issues rather than details, and the *efficiency* of the sector in the creation of equitable, stable, and sound business conditions under lower regulatory risk premiums.

The *yardstick idea* is practically implementable, compatible with the EC Directive 2009/72/EC and open for several types of cost functions. It leaves the difficult problem of asset valuation to the DSOs and the capital market. It also combines the firm's need for financial stability (*ex ante* tariff delegation) with the regulator's mission to ensure efficiency (*ex post* yardstick correction). The incentive parameters can be set to "tune" the regime to different capital risks.

The *quality incentive scheme* supports the optimal trade-off between cost and benefits of security of supply. Moreover, it provides quality incentive for DSOs irrespectively of their performances in the revenue yardstick competition. That is, even inefficient DSOs are encouraged to care about security of supply, investment analyses on quality provision can always be performed, irrespective of profit level.

The *proposal* is advanced in its use of mechanisms (revenue-based yardstick), yet the logic is intuitive and simple to explain to any stakeholder. Any Nordic customer in the NEMESYS model pays the lowest tariff that any comparable firm offers its clients. Any Nordic firm can define its profit as the difference between its costs and the lowest tariff charged by any other comparable firm. Comparability is defined on measurable dimensions of output, not accounting and process indicators. That's it.

References

Agrell, P. J., & Bogetoft, P. (2003). *Benchmarking for Regulation*. Report FP4, Norwegian Energy Directorate NVE, SUMICSID AB

Agrell, P. J., & Bogetoft, P. (2004). *Evolutionary regulation: From CPI-X towards contestability*. ENCORE position paper, University of Amsterdam

Agrell, P. J., & Bogetoft, P. (2005). *NVE network cost efficiency model*. Final report, Norwegian Energy Directorate NVE, SUMICSID AB

Agrell, P. J., & Bogetoft, P. (2007). *Development of benchmarking models for German electricity and gas distribution*. Final report AS6, Bundesnetzagentur, SUMICSID AB

Agrell, P. J., & Bogetoft, P. (2008). *Benchmarking the RTO Benchmarking Model in Norway*. Final report, EBL, SUMICSID AB

Agrell, P. J, Bogetoft, P., & Tind, J. (2005). DEA and dynamic yardstick competition in Scandinavian electricity distribution. *Journal of Productivity Analysis, 23*, 173–201

Bogetoft, P. (1994). Incentive efficient production frontiers: An agency perspective on DEA. *Management Science, 40*, 959–968

Bogetoft, P. (1997). DEA-based yardstick competition: The optimality of best practice regulation. *Annals of Operations Research, 73*, 277–298

Bogetoft, P., & Olesen, H. (2004). *Design of production contracts*. Copenhagen: CBS Press

CEER Council of European Energy Regulators (2001). *Quality of electricity supply: Initial benchmarking on actual levels, standards and regulatory strategies*. Autorità per l'energia elttrica e il gas

ECON (2000). *A model for estimation of expected energy not supplied*. Notat 59–00 (in Norwegian)

EMV (2004). *Guidelines for assessing reasonableness in pricing of electricity distribution network operations for 2005–2007*. Finnish Electricity Market Authority, 22 June 2004, Reg. no. 9/429/2004

EMV (2008). *Annual report 2007*. Finnish Electricity Market Authority

European Parliament and Council (2003). *Directive concerning common rules for the internal market in electricity and repealing Directive 96/92/EC*. (2003/54/EC) 26.06.2003

European Parliament and Council (2009). *Directive concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC*. (2009/72/EC) 13.07.2009.

European Commission (2008). *Attitudes of European citizens towards the environment*. Special Eurobarometer 295/Wave 68.2, March 2008

Førsund, F. R., & Edvardsen, D. F. (2001). *International benchmarking of distribution utilities*. Memo 35/2001, Frisch Institute, Oslo University

Littlechild, S. C. (1983). *Regulation of British telecommunication's profitability*. London

Majone, G. (1997). The agency model: The growth of regulation and regulatory institutions in the European Union. *EIPASCOPE 1997*(3):1–6

Meyer, M. A., & Vickers, J. (1997). Performance comparisons and dynamic incentives. *Journal of Political Economy, 105*(3), 547–581

Neeman, Z., & Orosel, G. O. (1999). *Contestable licensing*. Working Paper, Department of Economics, Boston University

NEMESYS (2005a). *Subproject A: System analysis, final report*. (Ed.) M. Syrjänen, SUMICSID AB and Nordenergi

NEMESYS (2005b). *Subproject B: Regulatory mechanism design, final report*. (Ed.) P. Bogetoft, SUMICSID AB and Nordenergi

NEMESYS (2005c). *Subproject C: Nordic efficiency model, final report*. (Ed.) H. Grønli, SUMICSID AB and Nordenergi

NEMESYS (2005d). *Towards a Pan-Nordic Regulation for Distribution System Operations, Final report*. (Ed.) P. J. Agrell, SUMICSID AB and Nordenergi

NVE (2007). *Endringer i forskrift 11. mars 1999 nr 302 om økonomisk og teknisk rapportering, inntektsramme for nettvirksomheten og tariffer*. Consultation document 18–2007 (in Norwegian)

SESSA (2005). *Harmonizing Effective Regulation: Scientific Consensus*. Summary paper from the Bergen Conference, European Regulation Forum on Electricity Reforms, Bergen, March 3–4, 2005

Shleifer, A. (1985). A theory of yardstick competition. *Rand Journal of Economics, 16*, 319–327

SOU (2007). *Förhandsprövning av nättariffer mm*. Interim report from the Energy network inquiry, SOU 2007:99

TNS Gallup, (2005a), TNS Gallups Energibarometer: God service betyr mer for strømkundene, *TNS Gallup Energy Newsletter*, April 2005 (in Norwegian)

TNS Gallup, (2005b), TNS Gallups Energibarometer: Strømkundene mangler kunnskap on nettleien, *TNS Gallup Energy Newsletter*, May 2005 (in Norwegian)

Williamson, O. E. (1976), Franchise bidding for natural monopolies: In general and with respect to CATV, *Bell Journal of Economics and Management Science, 7*, 73–104

Benchmarking in Regulation of Electricity Networks in Norway: An Overview

Endre Bjørndal, Mette Bjørndal, and Kari-Anne Fange

Abstract In this paper, we give an overview of the Norwegian regulation of electricity networks after the Energy Act of 1990 and the deregulation of the electricity markets in 1991. We concentrate on the regulatory oversight of distribution network companies and regional transmission. Our main focus is on the benchmarking models, including the application of their results, in the three periods of incentive regulation that we have seen so far, after its introduction in 1997. We examine the various data envelopment analysis (DEA) models that have been used, and we describe specific issues driving their development and how the results have been used.

1 Introduction

In Norway, the Energy Act came into force on January 1, 1991, and laid the foundation for market based production and power trading. Transmission and distribution were considered natural monopolies and remained regulated. The Norwegian electricity sector was deregulated, but never privatized, and the companies within the electricity sector are still, to a very large extent, under public ownership. An essential part of the restructuring of the industry was vertical separation of business activities exposed to competition and regulated operations, i.e., power production

E. Bjørndal (✉)
Department of Accounting, Auditing and Law, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: endre.bjorndal@nhh.no

M. Bjørndal
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Østfold University College, 1757 Halden, Norway
e-mail: mette.bjorndal@nhh.no

K.-A. Fange
Department of Business, Languages and Social Sciences, Østfold University College, 1757 Halden, Norway
e-mail: kari.a.fange@hiof.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_18,
© Springer-Verlag Berlin Heidelberg 2010

and trading on the one hand, and power transportation on the other hand. Statkraft, the major state owned power company, holding a large part of the national power production capacity, high voltage transmission network, and system operation, was split to form the generation company Statkraft, and Statnett, the system operator of the Norwegian power market, and owner of the main part of the transmission grid. Statkraft and Statnett are both owned by the Norwegian state. For other electricity companies vertical separation has been implemented by separation of accounts, and both regulated and non-regulated activities can be accomplished within the same companies. The competitive business segments include generation of power, power trading and retailing, and also for instance alarm services, broadband, and district heating. Network services and system operation are regulated by NVE, the Norwegian Water Resources and Energy Directorate.

In this paper, we consider the regulation of the distribution and regional transmission companies; we do not consider the regulatory oversight of the main transmission owner and system operator, Statnett. The main part of the paper describes the development of the benchmarking models that have been used since the introduction of incentive regulation in 1997, in order to determine efficiency requirements for individual companies when setting their allowed revenues. However, in the following section we first give a short description of the main elements of the regulation models for the three periods of incentive regulation that we have already seen, and then, after a general introduction to Data Envelopment Analysis (DEA), we treat some specific issues with regard to the DEA models that have been used by the Norwegian regulator, NVE.

2 Regulation of Electricity Networks After the Energy Act of 1990

During the first years after the Energy Act of 1990 and the starting point of the deregulation of the Norwegian electricity market, a rate of return (RoR) regulation was established from 1993. The main issues in this period were to determine book values of network assets in, for the most part, publicly owned firms, and an appropriate cost of capital. The former was determined to a large extent on the basis of new values, whereas the latter was established on the basis of a capital asset pricing model (CAPM) framework. But, already in 1997 a new regulation model was introduced, with more focus on providing incentives for cost efficiency in the development and operation of network assets and services.

The incentive regulation starting in 1997 has been based on total cost, since treating operating and capital costs differently may result in adverse incentives, as there are substitution possibilities among the two cost groups. Moreover, the incentive regulation has been implemented as a revenue cap, i.e., a maximum allowed revenue for individual companies. This is reasonable, since costs are mostly fixed, i.e., vary little with respect to transported volume, and demand for network services, which is a derived demand, is quite inelastic. Regulation by price caps was discussed before the regulation period starting in 2007 (e.g., von der Fehr et al. 2002), but was not

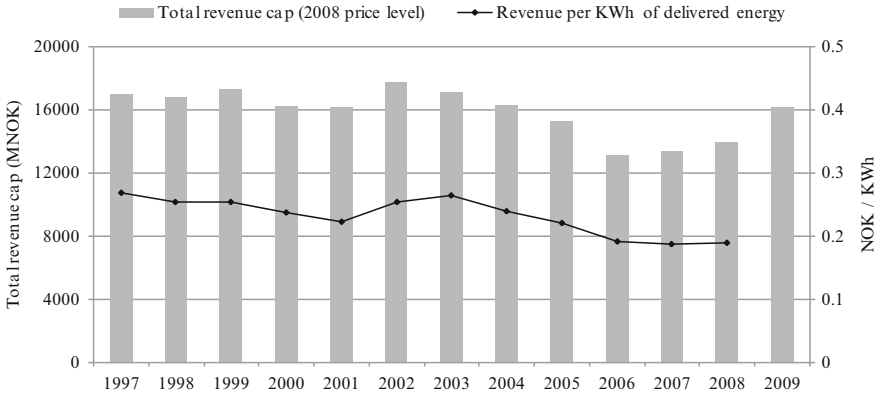


Fig. 1 Revenue caps for distribution and regional transmission

adopted. With fixed cost and inelastic demand, and assuming no ex post adjustments due to volume, the question of price or revenue caps is to a great extent a question of who is to bear volume risk, companies or customers. With a revenue cap it is the customers who bear the volume risk.

The regulation of the network companies is an ex ante regulation with some ex post adjustments, and the process is structured as follows. Before the period for which the allowed revenue is to be determined, cost data and information about company inputs and outputs are collected. This information is used to evaluate the relative efficiency of companies. Together with updates on prices, interest rates and possibly the level of activity of the companies' operations, total cost and efficiency results are used to settle on revenue caps for individual companies. Finally, after the period, when prices and cost are known, allowed revenues are adjusted. Adjustments may also be made due to limits on maximum and/or minimum accounting rates of return.

Since 1997, there have been three periods of incentive regulation:

- Period I: 1997–2001
- Period II: 2002–2006
- Period III: 2007–2011

In Fig. 1 we show some economic figures on aggregate level for the period 1997–2009. We show total revenue caps for distribution and regional transmission networks, and the corresponding revenue per unit of delivered energy (NOK/KWh).

In the following section, we describe some of the main elements of the Norwegian regulation of electricity distribution (mostly $\leq$22 kV) and regional transmission (mostly between 22 and 132 kV) during these three periods.

2.1 Period I: 1997–2001

In the first period of incentive regulation, total cost was based on accounting values from 1994 and 1995, and this was the starting point for calculating company specific

revenue caps for the whole period from 1997 until 2001. Within the period, the revenue cap was adjusted annually for inflation and changes in the energy prices (to account for power losses), for general and individual efficiency requirements and for increases in delivered energy in the concession area of the company. The revenue cap in 1997 was

$$Rcap_{1997} = C_{1997} \cdot (1 - X_g)$$

where C_{1997} is the actual cost from 1994 and 1995 adjusted to 1997 price levels, and X_g is a general efficiency requirement of 2%. The update of the revenue cap from year n to $n + 1$ was then accomplished through the following formula

$$Rcap_{n+1} = Rcap_{n,\,price\ adjusted} \cdot \left(1 + \frac{\Delta DE_{n,n+1}}{2}\right) \cdot (1 - X_{g+i})$$

where $\Delta DE_{n,n+1}$ is the percentage increase in energy delivered from one year to the next in the concession area of the company, and X_{g+i} is the sum of a general (1.5%) and individual efficiency requirement. The individual efficiency requirement was determined from DEA, and it was assumed that about half the inefficiency potential should be caught up with over the regulation period. The compensation for increases in delivered energy was introduced to account for necessary increases in cost due to new activities, and this constituted approximately 300 million NOK annually for the whole industry.

In 2001, a quality mechanism based on the value of lost load (VOLL) was introduced in the regulation. In this period, the quality mechanism represented an adjustment of the revenue caps, and this could be positive or negative. Expected VOLL was computed for each company, partly based on historical data and partly from model results. Any difference between actual and expected VOLL was charged or attributed to the companies.

Finally, minimum and maximum average accounting rates of return for the whole period applied, and they were set to 2% and 15%, respectively.

2.2 Period II: 2002–2006

The second period of incentive regulation was structured similarly to the first. Total cost was now based on accounting values from 1996 to 1999, and this formed the starting point for calculating annual revenue caps for 2002–2006. For 2002 the revenue cap was

$$Rcap_{2002} = C_{2002} \cdot (1 - X_{g+i})$$

where C_{2002} is total cost from 1996 to 1999 adjusted to 2002 price levels. Average operating cost and 1999 depreciation were adjusted by the consumer price index, while average network losses in MWh were evaluated at a reference price for energy for 2002 determined by NVE. Interest was calculated from depreciated book

values at the end of 1999, and with a regulated interest rate that in period II was updated annually. X_{g+i} was the sum of a general (1.5%) and individual efficiency requirement, the latter found by DEA, as in period I.

Within the period, the revenue cap was updated annually for inflation, changes in energy prices and interest rates, and for general and individual efficiency requirements. The revenue cap of year n in the 5-year regulation period was

$$Rcap_{2001+n} = C_{2001+n} \cdot (1 - X_{g+i})^n + CP_n$$

where C_{2001+n} is actual total cost from 1996 to 1999 adjusted to year $2001+n$ price levels, and CP_n is a compensation parameter for new investments. In regional transmission, the compensation parameter was based on actual new investments, whereas in distribution the compensation was accomplished by an index depending on new customers connected to the grid and the national increase in delivered energy. The compensation for new investments constituted approximately 200–300 million NOK annually for the whole industry in period II.

The quality mechanism was refined, but worked otherwise mostly as in period I, although, as we will describe later, it was also included in the benchmarking models. The minimum and maximum average accounting rates of return were 2% and 20%, respectively.

2.3 Period III: 2007–2011

Although the long time horizons of the first two regulation periods gave strong incentives for cost efficiency,[1] the same long time horizon had an adverse effect on investments. It took a long time before depreciation and interest for new assets were accounted for in total cost, and this could have severe effects on the net present values of new investments (Bjørndal and Johnsen 2004). In the third regulation period, we therefore saw some major changes, especially related to annual updates of cost and efficiency requirements, the latter taking the shape of cost norms from the DEA benchmarking. Thus, from 2007 annual revenue caps are established for individual companies based on a combination of actual cost and cost norms, according to the following yardstick formula:

$$Rcap = C + \rho(C^* - C) = \rho C^* + (1 - \rho)C, \tag{1}$$

where C is the actual cost, C^* is the cost norm, and $\rho \in [0, 1]$ is a factor that specifies the strength of the incentives in the yardstick model, i.e., the weight that is attributed to the cost norm.

[1] There could, however, also be ratchet effects, since total cost formed the basis for revenues in the *next* 5-year period.

For 2007 and 2008, ρ was equal to 0.5, but from 2009 onwards it has increased to 0.6. Actual cost and cost norms are updated annually, although, in practice, due to accounting procedures and the need for securing the quality of the data, up until now there has been a time lag in the application of cost data. More specifically, for 2007 and 2008 the cost data used for calculating actual cost and analyzing relative efficiency were 2 years old; i.e., the actual total company cost C estimated for year t consisted of a combination of registered and calculated costs, based on accounting values[2] in year $t-2$.

For distribution companies and regional transmission companies, the cost norm, C^*, is calculated based on relative efficiency scores found by DEA. There are still separate DEA models for distribution functions and regional transmission/central grid functions, respectively. A variant of super efficiency is implemented such that efficiency scores may be higher than 100%. When evaluating relative efficiency with DEA, average (industry) efficiency will depend on implementation details like, for instance, the number of evaluated companies (the size of the dataset), the number and specific choice of inputs and outputs, assumptions about scale efficiency, and whether or not super efficiency is modeled. In order to secure efficiency improvements over time and the attractiveness of the industry to investors and employees, it is important that particularly efficient companies can earn more than the normal rate of return. Thus, the efficiency scores are calibrated such that the representative company earns the normal rate of return. We discuss some of these DEA developments in the next section.

Due to the time lag in the use of accounting data, it was argued that new investments must be compensated in order to earn the normal rate of return in a representative company. This was accomplished through a compensation parameter, CP (this parameter and its use is discussed in Bjørndal et al. (2008b). The formula for establishing the revenue of a company in year t could then be written as:

$$Rcap_t = \rho C^{**}_{t-2} + (1-\rho)C_{t-2} + CP = \rho E^{*}_{t-2} C_{t-2} + (1-\rho)C_{t-2} + CP \quad (2)$$

where C_{t-2} is the price adjusted cost base from year $t-2$, E^{*}_{t-2} is the calibrated efficiency score of the company, and C^{**}_{t-2} is the corresponding calibrated cost norm. For the whole industry, the value of the compensation parameter has been calculated to 300–400 million NOK. From 2009, the time lags have been removed, so that there is no longer need for the compensation parameter.

[2] Operating and maintenance costs from the year $t-2$ were adjusted for inflation, depreciation set equal to the accounting values in the year $t-2$, while network losses (NL) were found by taking the losses in MWh in the year $t-2$ and multiplying by an average area price (based on Nord Pool Spot) for the year t. The cost of capital was found by multiplying the book value (BV) of the company assets at 31.12 in the year $t-2$ by the NVE rate of return, r_{NVE}, for the year t. This regulated rate of return is determined annually, based on a risk free rate of return and a risk premium. Finally, total cost includes the value of lost load (VOLL) which is calculated as lost load times a unit price, with different unit prices for various customer groups.

3 Benchmarking and Productivity Measurement for Regulation

In order to establish reasonable revenue caps for network companies under incentive regulation, it is necessary to analyze company performance. We distinguish between analysis of productivity (absolute performance) and efficiency (relative performance), and most of the analyses performed for regulation purposes belong to the latter category. The two most widely used methods for efficiency analysis are Stochastic Frontier Analysis (SFA) and Data Envelopment Analysis (DEA). The former method belongs to the group of so-called parametric methods, where one assumes a general mathematical function for the relationship between inputs and outputs, and estimates its parameters. On the other hand, DEA takes a non-parametric approach, whereby the efficient frontier is fitted directly to the data. Hence, under DEA there is no need to assume a particular mathematical structure, although one still has to make choices with respect to the assumptions that define the set of feasible production plans and the efficient frontier. There are also alternatives to SFA and DEA, such as Corrected Ordinary Least Square method (COLS),[3] and Stochastic Data Envelopment Analysis (SDEA). An introduction to different benchmarking methods is given by Coelli et al. (2005).

The Norwegian regulator has mainly used DEA for its efficiency analyses. In this section we give a brief introduction to DEA and discuss some important implementation issues related to the Norwegian regulation regime.

3.1 Introduction to DEA

Figure 2 illustrates the computation of technical efficiency under DEA. It shows an example with one product (y) and one input factor (x), and with five companies A–E. The figure illustrates some of the basic assumptions that are commonly used in DEA and that defines the set of feasible production plans and the efficient frontier. First, note that all DEA models assume that the observed data belongs to the production possibility area, i.e., that they correspond to feasible production plans. It is also common to assume that "synthetic" companies can be constructed by taking convex combinations of the existing data points, as illustrated by the area ABCE in the figure. With production possibility area ABCE, the efficient frontier consists of line segments ABC, i.e., input/output combinations such that output cannot be increased without also increasing input. Another common assumption is that of free disposability, i.e., that surplus quantities of input and output can be disposed of without cost. The latter assumption implies that if (x, y) is a feasible production plan, then (x', y') will also be feasible if $x' \geq x$ and $y' \leq y$. This gives the extension of the production possibility area indicated by horizontal lines in Fig. 2. Convexity

[3] COLS estimates model parameters using OLS and shifts the intercept of the regression line such that it passes through the minimum observation (Lowry and Getachew 2009).

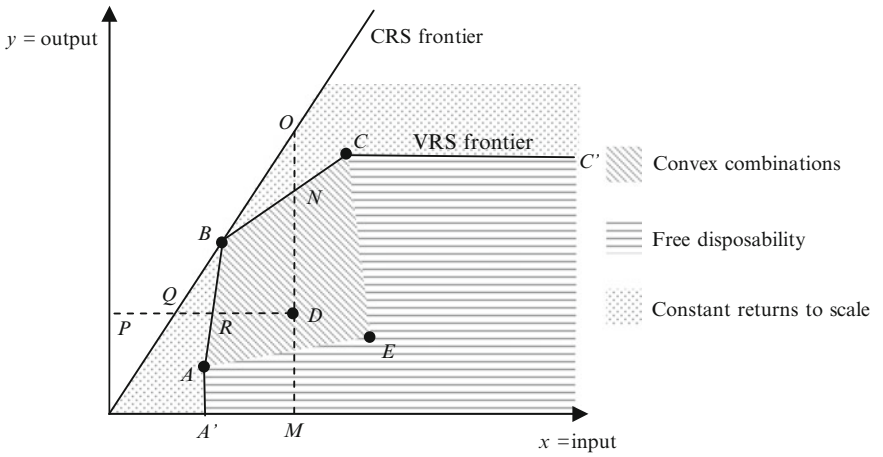


Fig. 2 Technical efficiency under Constant Returns to Scale (CRS) and Variable Returns to Scale (VRS)

and free disposability give rise to the so-called Variable Returns to Scale (VRS) technology, with a corresponding efficient frontier given by A'ABCC' in Fig. 2.

Once we have determined the efficient frontier, the efficiency of a particular company can be evaluated by comparing the corresponding data point to a reference point on the frontier. Suppose we wish to evaluate company D and that we are willing to accept the VRS frontier as the correct one. Note that we have an infinite number of reference points to choose from, since it is not obvious in which direction we should move from point D to the VRS frontier. Figure 2 illustrates two typical choices with respect to direction, namely the horizontal (input) and vertical (output) direction. If we choose the input direction, as in the DEA models used for the Norwegian network companies, the reference point for company D will be point R, and the efficiency score can be computed as the ratio PR/PD. On the other hand, if the output direction is chosen, the efficiency score of company D is given by the ratio MN/MD. The input efficiency score PR/PD indicates the potential for reduction in input usage by company D, while the output efficiency score focuses on the potential for increasing output.

In some settings it is also reasonable to assume that any feasible production plan can be freely scaled up or down; i.e., if (x, y) belongs to the production possibility set, this is true also for (tx, ty), where t is a non-negative constant. This extends the production possibility area by the dotted area in the figure. The resulting production technology is commonly referred to as a Constant Returns to Scale (CRS) technology. The CRS efficient frontier is the straight solid line going through the origin in Fig. 2, and the efficiency score of company D can be computed as the ratio PQ/PD or MO/MD, depending on which direction one chooses towards the CRS frontier. Note that, since the frontier is a straight line that passes through the origin, we will have PQ/PD = MD/MO; hence, the input efficiency score can be found as the inverse of

the output efficiency score. This illustrates a general property of CRS models, i.e., it does not matter whether we choose the input or output direction when we evaluate the efficiency of a company. We also note that the CRS frontier lies further away from the observed data points than the VRS frontier, with the exception of the tangency point B. Therefore, CRS input (output) efficiency scores will always be lower (higher) than, or equal to, the corresponding VRS scores.

The technical efficiency measure discussed above evaluates the potential for input (output) reduction (increase). In the models that have been used by NVE, input use is measured in terms of cost. DEA cost efficiency models not only measure the potential for reduction in input usage, but also the potential for cost reductions through reallocation between input factors. An illustration of how this can be measured is given in Fig. 3. In this example, four companies produce the same quantity of output using two inputs, x_1 and x_2. Company D is faced with the factor prices w_1 and w_2, determining the slope of its isocost line. The optimal plan for company D would thus be to choose the same input mix as company B, and the overall cost efficiency of D can be expressed as *OP/OD*. Technical efficiency measures the distance from point *D* to the efficient frontier A'ABCC', i.e., the ratio *OQ/OD*. Allocative efficiency measures the additional cost reduction by improving the input mix at given prices, and can be expressed as *OP/OQ*. The overall cost efficiency can thus be decomposed into technical and allocative efficiency in the following way:

$$\frac{OP}{OD} = \frac{OP}{OQ} \cdot \frac{OQ}{OD}$$

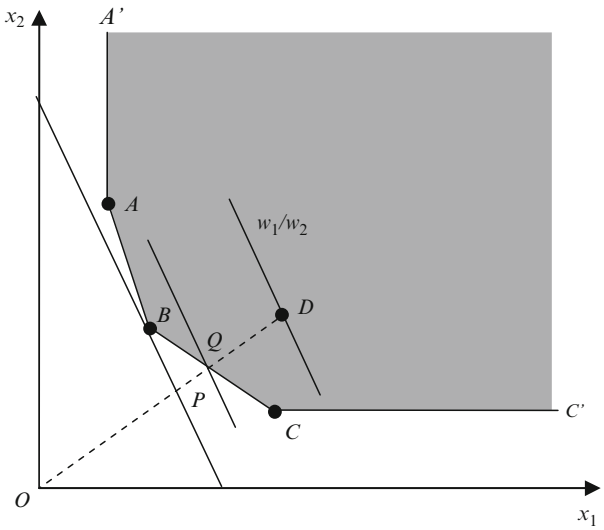


Fig. 3 Technical efficiency and cost efficiency

In other words, overall cost efficiency is equal to the product of allocative and technical efficiency. For a more comprehensive introduction to the DEA methodology, see Cooper et al. (2007).

3.2 Model Specification and Data Measurement Issues

Equations (3)–(7) below describe a DEA cost efficiency model. The variable x_{ij} represents company j's use of input-factor i, $i = 1, \ldots, m$, while y_{rj} represents company j's production of output r, $r = 1, \ldots, s$. In order to evaluate the cost efficiency of a particular company j, we use the factor price w_{ij} observed by that company for each input i that the company uses. The index j_0 represents the company that we want to evaluate. The decision variable z_i represents the optimal use of input i for the evaluated company. Hence, the objective function (3) measures the ratio between the optimal cost (cost norm) of the evaluated company, and its actual cost. The optimal cost corresponds to a set of peers, and the decision variable λ_j measures the weight of company j in this reference set. Equation set (4) requires the optimal input quantities of the reference company to be no greater than the optimal input quantities of the evaluated company, and equation set (5) requires the output quantities of the reference company to be at least as great as the corresponding quantities for the evaluated company. Equation (6) enforces the VRS restriction, while (7) ensures non-negative weights of the reference companies.

$$\underset{\lambda,z}{\text{Min}} \frac{\sum_i w_{ij_0} z_i}{\sum_i w_{ij_0} x_{ij_0}} \tag{3}$$

subject to

$$z_i \geq \sum_j \lambda_j x_{ij} \qquad i = 1, \ldots, m \tag{4}$$

$$y_{rj_0} \leq \sum_j \lambda_j y_{rj} \qquad r = 1, \ldots, s \tag{5}$$

$$\sum_j \lambda_j = 1 \tag{6}$$

$$\lambda_j \geq 0 \qquad j = 1, \ldots, n \tag{7}$$

If we can assume that all the inputs have strictly positive factor prices, we can replace (4) by an equality, and substitute the expression for z_i in the objective function; i.e., we can replace (3) and (4) by the following objective function:

$$\underset{\lambda}{\text{Min}} \frac{\sum_i w_{ij_0} \sum_j \lambda_j x_{ij}}{\sum_i w_{ij_0} x_{ij_0}} \tag{8}$$

The input-oriented VRS model given by (3)–(7), or equivalently, (5)–(8), was used in the regulation of the Norwegian network companies from 1997 to 2006. In the new regulation regime that was implemented from 2007, two important changes were made to the underlying DEA model. The CRS assumption, illustrated in Fig. 2, was introduced. Mathematically, this is equivalent to dropping restriction (6) in the LP-problem. In addition, the regulator decided to go from a model with five input factors, with corresponding factor prices, to a model with only one input, namely total cost, with a factor price equal to one. Mathematically, these two changes result in the following model:

$$\underset{\lambda}{\text{Min}} \frac{\sum_j \lambda_j x_j}{x_{j_0}} \tag{9}$$

subject to

$$y_{rj_0} \leq \sum_j \lambda_j y_{rj} \qquad r = 1, \ldots, s \tag{10}$$

$$\lambda_j \geq 0 \qquad j = 1, \ldots, n \tag{11}$$

where x_j is the total cost of company j. The numerator in (9) thus gives the value of the optimal cost norm of the evaluated company j_0, while the denominator is the actual cost of the company. Note that since the denominator is a constant, the optimal solution with respect to the λ's will not change if we replace (9) by the following expression:

$$\underset{\lambda}{\text{Min}} \sum_j \lambda_j x_j \tag{12}$$

By using (12) we obtain the cost norm for the evaluated company directly as the value of the objective function. The shadow prices of the output restrictions (10) will now be expressed in monetary units, making them easier to interpret.

The shadow prices can also be obtained by solving the dual to (10)–(12), given by:

$$\underset{p}{\text{Max}} \sum_r p_r y_{rj_0} \tag{13}$$

subject to

$$\sum_r y_{rj} p_r \leq x_j \qquad j = 1, \ldots, n \tag{14}$$

$$p_r \geq 0 \qquad r = 1, \ldots, s \tag{15}$$

where p_r is the price of output r. The dual LP-problem has an interesting interpretation, in which company j_0 optimizes non-negative prices for its outputs such that the resulting revenue, given by the value of (12), is maximized. The choice of prices is, however, restricted by (14), saying that no company, including company j_0, can have positive profit when evaluated at these output prices.

In the rest of this section we discuss data and model specification issues with respect to the DEA models that have been used as part of the Norwegian regulatory regime since 1997.

3.2.1 The Number of Input Factors

Since the outputs of an electricity network company are mostly outside of the company's control, it makes sense to use an input-oriented DEA model with cost as input(s), and where the output factors are assumed to be exogenously given. The input factors that have been used are shown in Table 1.

The models in regulation periods I and II had the same input set, except that a quality cost variable (value of lost load) was added in regulation period II, increasing the number of inputs from four to five. In periods I and II, there were two versions with respect to capital costs, one based on reported book values and the other based on a catalogue of standard values. Separate DEA analyses were performed for each of the capital definitions, and the final efficiency score for a company was set equal to the maximum of the two efficiency scores.

From period III it was decided to switch to a model with only one input, i.e., as given by (9)–(11), and to use only book values as basis for evaluating capital costs. The five elements that constitute total cost are shown in Table 1. In the rest of this section we discuss the former change, while capital costs is the subject of the next section.

It is easy to show that if two input factors i and k have identical factor prices, i.e., if $w_{ij} = w_{kj}$ for all companies, then we can replace them by a single input l, defined as $x_{lj} = x_{ij} + x_{kj}$, without changing the efficiency score given by the value of (8). Table 1 shows that this applies to goods & services and VOLL. In the case of power losses there was a common factor price for all companies, given by the average system price. By rescaling the quantities and prices for this input factor,

Table 1 Input factors in the various DEA models

Variable	Period			Unit of measurement	Factor price
	I	II	III		
Labor	x	x	x	No. of man-years	Company-specific average wage
Capital, book values	x	x	x	NOK	Depreciation factor + r_{NVE}
Capital, catalogue values	x	x		NOK	Annuity factor based on r_{NVE} and observed asset lifetimes
Goods & services	x	x	x	NOK	1
Power losses	x	x	x	MWh	Based on the system price of power from Nord Pool Spot
Value of lost load (VOLL)		x	x	NOK	1

we can change the factor price of losses to unity without altering the value of (8). Hence, it would have been possible to reduce the number of inputs in regulation period II from five to three without changing the results.

For the remaining two inputs, labor and capital, there was considerable variation among the companies. Figure 4 shows the distribution of the wage numbers in the dataset for period II. Since a company's factor price for labor is the *average* wage for its employees, we would expect to see moderate variations among the companies. The large variations that we see in the figure suggest the existence of reporting errors, see the discussion in Bjørndal et al. (2004).

Figure 5(a) illustrates the effect of variation in the factor prices. We have replaced the individual factor prices with, for each input factor, a weighted average over all the companies; i.e., we measure efficiency assuming all companies have access to the same input factor market with a common factor price. We see that the effect of this operation is quite small, except for one company, corresponding to the outlier in Fig. 4.

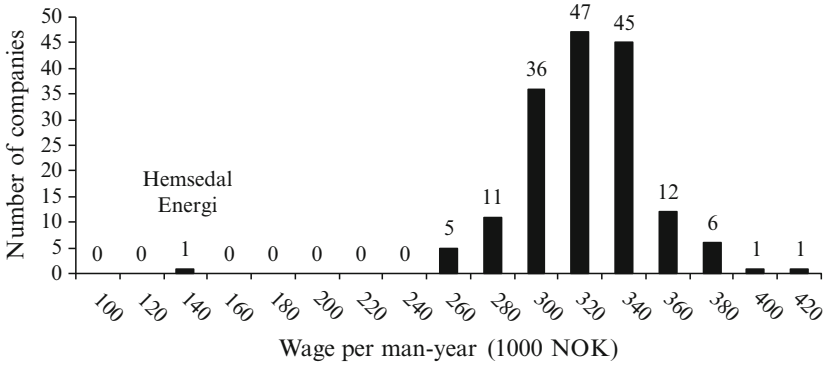


Fig. 4 Wage factor prices in the distribution network dataset for 1996–1999

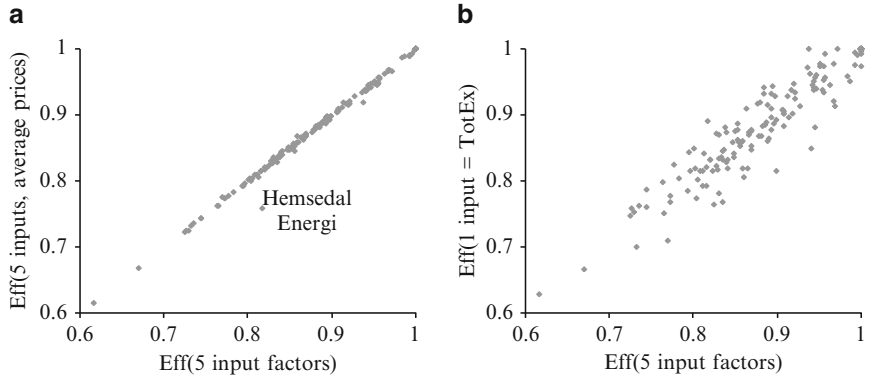


Fig. 5 Effect of (**a**) using average instead of individual factor prices and (**b**) replacing the five inputs with a single input factor (TotEx), 1996–1999 dataset

Table 2 Quantity versus price effects in the case of labor expenses

Case	Number of man-years	Wage per man-year	Total wage expenses	Efficiency score (%)	Change-(%)
Initial	94.05	311,112	29,260,084	87.96	
I	94.05	50,000	4,702,500	90.63	2.67
	15.12	311,112		100.00	12.04
II	94.05	250,000	23,512,500	88.59	0.63
	75.58	311,112		91.02	3.06
III	94.05	500,000	47,025,000	86.25	−1.71
	151.15	311,112		79.58	−8.38

It is tempting to conclude from the analysis in Fig. 5(a) that, since measuring efficiency with company-specific factor prices gives almost the same results as when using a common factor price, it should be possible to get rid of the remaining two input factors as well, with very small changes in the efficiency results. This is, however, not correct, as illustrated by Fig. 5(b), where we compare the results from model (5)–(8) and model (9)–(11), with five and one input(s), respectively. In order to understand this apparent puzzle, note that rescaling the labor and capital factor prices so that they are equal to unity for all companies also requires rescaling of the corresponding factor *quantities*. As it turns out, changes in factor prices and factor quantities can have very different effects on individual efficiency scores, even if they result in the same change in total cost. This is illustrated by the example in Table 2, which refers to a company with initial labor expenses of 29.3 MNOK and an efficiency score of 87.96%. Case I shows what would happen if labor expenses were reduced to 4.7 MNOK. If this is done via a reduction in the factor price, i.e., from 311,112 to 50,000 NOK, the efficiency score would increase by 2.67%. The same cost reduction could be obtained by reducing the number of man-years from 94.05 to 15.12, as shown in the table. In this case, however, we see a much larger increase in the efficiency score, by 12.04%. Cases II and III further illustrates that changes in factor quantities have a much stronger effect on efficiency scores than equivalent changes in factor prices.

More details about these examples can be found in Bjørndal et al. (2004). For a general treatment of the choice of input factors, see Dyson et al. (2001). There is also some literature on input aggregation, see e.g., Tauer (2001) and Färe et al. (2004).

3.2.2 Capital Costs and Age Effects

The electricity network industry can be characterized as capital intensive, with large investments in equipment with long asset lifetimes. Therefore, the quality of the efficiency analyses depends heavily on the way capital costs are measured. Alternative methods for calculating capital costs exist, and in practice the choice is often between methods based on linear depreciation or annuity-based methods. In the first two regulation periods in Norway, both of these models were used side by side, as

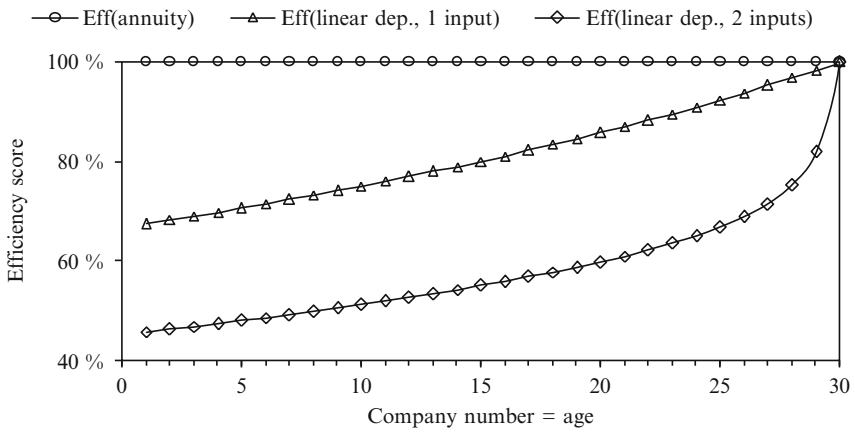


Fig. 6 Efficiency scores for an industry consisting of 30 vintages of a representative company

shown in Table 1. In the third regulation period only one method, with linear depreciation based on book values, is used.

From industry representatives it is often claimed that equipment productivity is nearly constant throughout the life span of the equipment, which suggests the use of annuity-based methods. In fact, using book values and linear depreciation may lead to a negative bias in the efficiency scores, as illustrated by the stylized example in Fig. 6. Here, we have created a dataset with 30 companies, where the only difference between the companies is their age. The companies are assumed to have two types of costs, capital costs and operating costs.[4] The figure illustrates that efficiency analysis with annuity-based capital costs yields efficiency scores that are independent of age, whereas efficiency scores based on linear depreciation will be increasing with respect to age. In the latter case, the efficiency scores will also depend on whether we use model (5)–(8) with two inputs or model (9)–(11) with total cost as the single input, since factor prices in this case will differ among companies.[5]

Looking at the dataset that was used in the second regulation period, we find some evidence of age bias, as illustrated in Fig. 7. The diagram shows that efficiency scores are significantly lower for "young" companies, i.e., where the ratio between book value and catalogue value (new value) is relatively high.

The efficiency analyses are used to compute the cost norms for the regulation model; hence, a bias in the efficiency scores may influence the revenue caps. In Fig. 8 we illustrate this phenomenon using the same example as in Fig. 6. We assume

[4] The interest rate is 5%, and the operating costs have been set equal to the annuity-based capital costs.

[5] In the cost efficiency model used in the first two regulation periods, the factor price for capital in the book value model was set equal to the sum of the depreciation rate and the interest rate, where the depreciation rate was calculated as depreciation divided by book value. Since book values decrease with age, and depreciation is constant, factor prices will differ due to age.

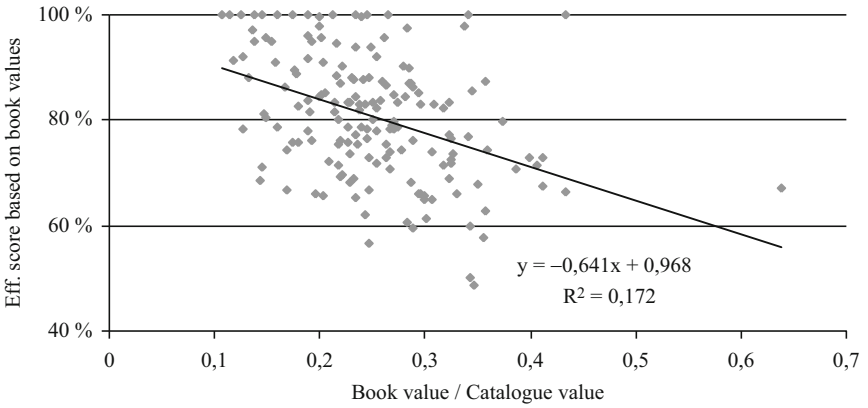


Fig. 7 Efficiency scores versus age in the 1996–1999 dataset for distribution companies

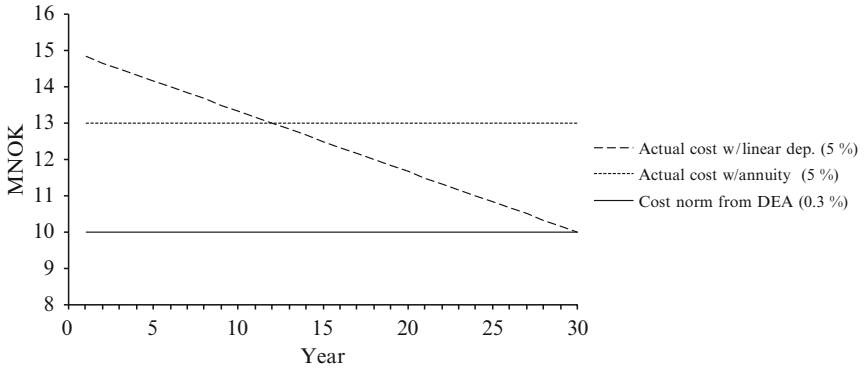


Fig. 8 Actual costs and cost norms over the life span of a company

that the companies last for 30 years, and that one company is added to/removed from the dataset each year. Apart from the ages of the existing companies, there are no changes to the industry over time. The diagram illustrates the development in cost and cost norm for a company during its 30-year life span. The dashed and dotted lines in the figure show calculated costs based on linear depreciation and annuities, respectively. We assume that the revenues are set equal to the cost norms; hence the dashed and dotted lines correspond to revenue caps under a rate of return regulation regime. Both of these alternatives have a profitability (IRR shown in parentheses) equal to the cost of capital. If, however, we let the revenue caps be determined by a cost norm based on DEA analysis with book values, the revenue level will be set equal to the cost of the oldest company in the dataset, i.e., a 30-year-old company. If the age bias is not compensated for, the resulting profitability will be only 0.3%.

The obvious way to correct the age bias would be to define the capital costs to be constant over time, e.g., by using annuities and catalogue values. If this is not viable, e.g., due to lack of data, there are several alternatives. One alternative is to introduce an age parameter as an extra output parameter in order to correct the bias. Alternatively, one could correct the bias by adjusting the efficiency scores/cost norms *after* running the DEA analysis. Such a calibration of the efficiency scores could be implemented for a number of reasons, not only age bias, and will be discussed in one of the sections below. Both alternatives are discussed in detail by Bjørndal and Bjørndal (2006a, b) and Bjørndal et al. (2008b). The use of an age parameter is related to the discussion of *environmental* variables in Dyson et al. (2001).

3.2.3 Choice of Output Variables

We can distinguish between output variables that describe characteristics of the companies themselves versus variables that serve to describe the environment in which the companies operate. Some of the variables that are listed in Tables 3 and 4 below clearly belong to the first category, such as delivered energy and the number of customers, while others, such as forest, snow and coast, belong to the latter. Network size variables such as HV and LV lines cannot be easily classified as either "pure output" or "environmental" variables. The motivation behind their inclusion is to represent demographical and topological factors that influence the companies' network size and cost level, and it is lack of available data that has made it necessary to represent these factors using input variables as proxies.

Note that some of the environmental factors are represented by indices. For instance, the forest index is given as a percentage value, measuring how much of a concession area is covered by high-growth forest. Indices must be correctly scaled

Table 3 Output variables for distribution networks

		Regulation period		
Variable	Unit of measurement	I	II	III
Delivered energy	MWh	x	x	x
Customers	No. of customers	x	x	
Customers, except cottages	No. of customers			x
Customers, cottages	No. of customers			x
HV lines	Kilometers	x	x	x
LV lines	Kilometers	x	x	
Sea cables	Kilometers	x		
Expected VOLL	NOK		x	
Network stations	No. of stations			x
Interface	Weighted measure			x
Forest	Forest index × HV overhead lines			x
Snow	Snow index × HV overhead lines			x
Coast	Coast index × HV overhead lines			x

Table 4 Output variables for regional transmission networks

Variable	Unit of measurement	Regulation period		
		I	II	III
Transported effect	MW	x	x	
Network size	Weighted value	x	x	
Exchange	Weighted value	x	x	
Central grid tasks	Weighted value	x	x	
Expected VOLL	NOK		x	
Lines, air	Weighted value			x
Lines, earth	Weighted value			x
Lines, sea	Weighted value			x
Interface	Weighted value			x
Forest	Forest index × Overhead lines			x

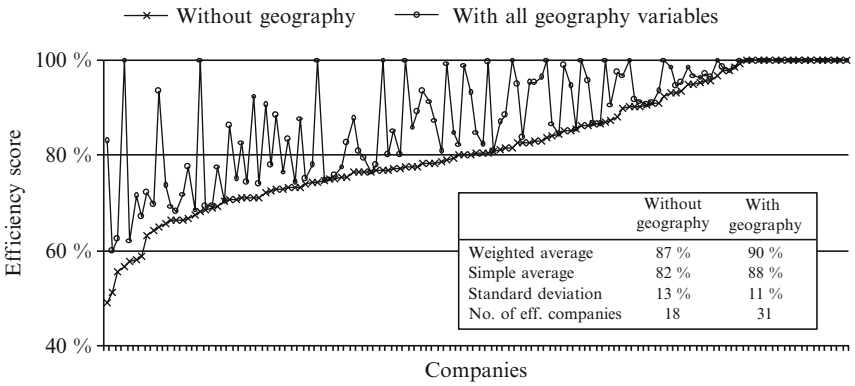


Fig. 9 Effect of adding geography variables (forest, snow, coast), 2006 dataset

before used together with scale-dependent variables in a DEA model, otherwise the results will be biased in favor of small companies.[6]

Note that, because of the output restrictions given by (5), adding a new variable to the output set will have a non-negative effect on the efficiency scores. An example is shown in Fig. 9, which illustrates the effect of adding the so-called "geography" variables to the DEA model for distribution networks in regulation period III. We see that the new variables cause an increase in the average efficiency score from 87 to 90%. Since the maximum obtainable efficiency score is 100% and the number of efficient companies can only increase when we add variables, the dispersion of the efficiency scores will inevitably decrease. This is a well-known problem, i.e., adding more variables will reduce the discriminatory power of the DEA model, and one should therefore try to restrict the number of variables relative to the number

[6] See Dyson et al. (2001).

Table 5 Test statistics for HV-lines and LV-lines, pooled dataset for 2001–2004

	Already included variables					
New variable	Customer	Energy	HV	LV	t	$p(T > t)$
HV	x	x			17.3	0.00
LV	x	x			8.7	0.00
HV	x	x		x	9.4	0.00
LV	x	x	x		1.4	0.08

of companies in the dataset. As a rule of thumb, Dyson et al. (2001), for instance, recommend that the number of companies should be at least $2m \times s$, where m and s are the number of input and output variables, respectively.

In order to evaluate potential variables for inclusion in the output set, there exists statistical tests (see Banker and Natarajan 2004). Such tests have been used in the development of the various DEA models that are discussed here (see Kittelsen 1993 and NVE 2006a). In order to illustrate some implications of the testing procedures, we use an example of a test that was done in connection with the development of the third regulation model. Table 5 shows four different T-tests that tests whether the mean efficiency score increases significantly (with a 5% significance level) as a result of adding either HV or LV as a new output variable.[7] We see that both variables have a significant effect if they are added to an output set consisting of the customer and energy variables. However, if the LV variable is added to an output set already containing the HV variable, it does not have a significant effect. The opposite is not true, i.e., the HV variable has a significant effect even though the LV variable is already in the output set. The example shows that the outcome of the test procedures may indeed depend on the sequence in which the variables are tested. In NVE (2006a), the HV variable was included in the output set from the start, i.e., it was never tested, and the conclusion was therefore that the effect of the LV variable was not significant. The regulator used this, combined with suspected low data quality, as an argument for dropping the variable. The regulator's decision was criticized,[8] and the main counter-argument was that HV and LV are both endogenous variables. By including one of them, while at the same time excluding the other, we will distort the investment incentives of the companies. It could, therefore, be argued that the LV variable should be included in the output set even though it does not pass the statistical test.

3.2.4 Scale Assumption

Assumptions with respect to economies of scale can have a significant effect on the efficiency evaluation of individual companies, as illustrated by Fig. 2, where

[7] The formulas can be found in Kittelsen (1993).

[8] See NVE (2006b).

Table 6 The largest company with respect to each output factor, 1996–1999 dataset

Output	No. of units	Company	No. of customers
Low voltage lines (km)	8,951	BKK Distribusjon AS	147,500
High voltage lines (km)	4,969	Nord-Trøndelag El.nett	73,557
Customers	303,312	Viken Energinett AS	303,312
Delivered energy (MWh)	8,370,400	Viken Energinett AS	303,312
Expected VOLL (1,000 NOK)	48,089	Troms Kraft Nett AS	59,376

we see that the CRS frontier results in a much stricter evaluation of small and large companies than the VRS frontier. The VRS restriction (6), combined with the output restriction (5), implies that for VRS models a company that is largest with respect to an output is automatically 100% efficient, since it must be its own reference. To illustrate the consequences of this property for the regulation, we show in Table 6 the largest companies with respect to output parameters in the distribution dataset from regulation period II. Since there were four companies that were evaluated as 100% efficient irrespective of their cost level, we claim that a large share of the industry, representing some 600,000 out of 2.5 million customers, was rather weakly regulated. In fact, the weak regulation of large companies with VRS models was stated as one of the main reasons for switching to a CRS model in regulation period III (see NVE 2006a). The DEA literature also proposes statistical tests for determining the appropriate assumptions with respect to economies of scale (see Banker and Natarajan 2004).

3.2.5 Super Efficiency and Incentives

A regulation scheme should give the regulated companies strong incentives for cost efficient investment and operating decisions. This implies that a company's cost norm should be independent of its actual cost. This is especially apparent in the yardstick regulation regime that was introduced in 2007, where a new cost norm is established each year via the efficiency analyses. When a company reduces its cost level, this should not lead to a reduction in the cost norm, since that would give the company weaker incentives for cost reductions. This property is not fulfilled for a 100% efficient company, however, since the cost norm for such a company will be set equal to its actual cost.

One way to avoid this phenomenon is to apply the procedure suggested by Andersen and Petersen (1993), whereby the evaluated company is excluded from the dataset. We see an example of this in Fig. 10. The revised efficiency scores are only different for those companies that would otherwise have an efficiency score of 100%. Some companies would get very high efficiency scores if this procedure was to be used, and the regulator chose a modified procedure, as described in NVE (2006a). According to the revised procedure, super efficient companies are re-evaluated against a dataset from the year(s) preceding the year of the current

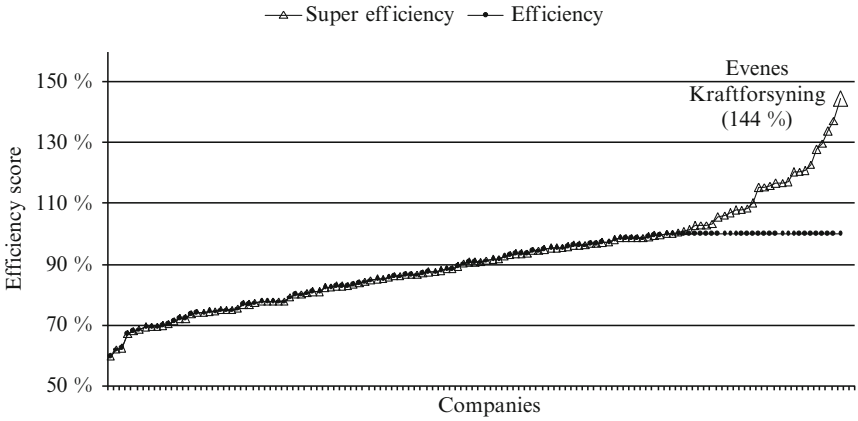


Fig. 10 Ordinary efficiency and super efficiency scores, 2006 data for distribution companies

Table 7 Incentive effects of DEA models with (1) no super efficiency, (2) unrestricted super efficiency, or (3) restricted super efficiency á la NVE

	Model/Year (t)	20 x 1	20 x 2	20 x 3
ΔC_t		−10	−10	−10
ΔC_t^*	1	−10	−10	−10
	2	0	0	0
	3	0	−10	−10
$\Delta Rcap_t$	1	−10	−10	−10
	2	−4	−4	−4
	3	−4	−10	−10
$\Delta \Pi_t$	1	0	0	0
	2	+6	+6	+6
	3	+6	0	0

dataset. The DEA model in the second step includes data for the company itself; hence, a company can only appear as super efficient if it has improved its performance relative to the previous year(s).[9]

Table 7 illustrates the incentive effects of the different models with respect to super efficiency, given a yardstick regulation model as in (1), with $\rho = 0.6$. Suppose a company with an efficiency score of 100% is considering an investment decision that will result in a permanent reduction in the annual cost of 10 MNOK. If the DEA model does not consider super efficiency at all (Model 1), the cost norm of the company will be set equal to its actual cost; hence, the reduction in the revenue cap will

[9] For 2007, the initial DEA analyses were based on data from 2005, and the re-evaluation in the second step was against a dataset from 2004. For 2008 and 2009, the re-evaluation dataset was created by taking averages of the datasets for the periods 2004–2005 and 2004–2006, respectively.

be equal to the cost reduction, and the company will have no incentives to reduce its cost level. Model 2 measures super efficiency á la Andersen and Petersen (1993); hence, the cost norm will be unaffected by the change in the actual cost level. With this model, therefore, the company will experience a permanent increase in its profit equal to 60% of the cost reduction. Model 3 corresponds to the "restricted" super efficiency measure that is used by the regulator, where the company's efficiency is re-evaluated against a dataset containing its own cost and output data for the previous year. In the first year the cost norm will be set equal to the company's cost for the previous year; hence there will be no reduction in the cost norm, and the profit for that year will be the same as with Model 1. From the second year and onward, the cost norm will catch up with the actual cost level, and the effect on profits will be zero. Although this example is somewhat simplified,[10] it clearly illustrates that a model where little or no super efficiency is allowed will provide relatively weak incentives for cost reductions.

3.2.6 Calibration and Average Profitability

The discussion of super efficiency in the previous section shows that the *marginal* incentives in the industry depend on the specification of the benchmarking model. However, model specification also influences the *average* profitability of investments/companies, as illustrated by Table 8 below, where we show average and cost weighted average efficiency for various versions of the DEA model (scale assumption, super efficiency). From the cost weighted average efficiency score we can also compute the total cost norm for the whole industry. In the present yardstick model, only companies with an efficiency score of at least 100% will be able to earn the NVE rate of return. Since the NVE rate of return is determined as the sum of the risk free rate and a suitable risk premium estimated using CAPM, it is reasonable that some companies, if they manage to run their business in an efficient manner, should be able to earn *more* than the NVE rate of return. This may be achieved to some extent by introducing super efficiency in the model. However, as can be seen from Table 8, the industry cost norm will be only 91% of the actual industry cost

Table 8 Industry average of the efficiency scores for various DEA models. Distribution networks, 1996–1999 dataset

	VRS	CRS	CRS w/super efficiency	Modified super efficiency (NVE)
Simple average (%)	88	85	88	85
Cost weighted average (%)	93	88	91	89
Industry cost norm (MNOK)	9,168	8,666	8,948	8,709

[10] Time lags are not considered. Also, the regulator has gradually increased the number of years that the dataset in the second step is based on, cf. footnote 9.

if super efficiency is introduced, compared to 88% for the comparable CRS model without super efficiency.

This raises an interesting question, namely, how much of the total cost should the industry as a whole be allowed to collect in the form of revenues? It does not seem fair that the total revenue level of the industry should be determined by somewhat arbitrary model specification choices, as illustrated by Table 8. In the first two regulation periods, the initial revenue level was set equal to the companies' actual costs, with subsequent reductions given by general and individual efficiency requirements. For the third regulation period, the regulator decided to calibrate the cost norms such that the industry revenue would be set equal to the sum of actual costs for the industry.[11]

There are, of course, a number of ways by which the initial revenue shortfall could be distributed among the companies, and the following three methods have been used by the regulator so far in regulation period III:

(a) To normalize the efficiency scores such that the cost weighted average becomes equal to 100%. This is equivalent to distributing the revenue shortfall in proportion to the initial cost norms of the companies.[12]
(b) To distribute the revenue shortfall among the companies in proportion to their capital values.
(c) To add a constant to the efficiency score of each company, such that the cost weighted average becomes equal to 100%. This is equivalent to distributing the revenue shortfall in proportion to the actual costs of the companies.[12]

The methods differ with respect to the effect on *marginal* incentives and *average* profitability. To see the difference in marginal incentives, note that the basis for distributing the revenue shortfall is different. Actual capital and cost values depend on decisions taken by the companies, while cost norms, at least in principle, cannot be influenced by the companies' own decisions. This gives method A an advantage over the other two methods as far as incentives are concerned. The differences with respect to average profitability follows from the different cash flow time profiles, as illustrated by Fig. 11 for the same example that we used in Fig. 8. The lines marked with circles, squares and triangles correspond to calibrated cost norm profiles for the respective calibration methods. Method A represents a vertical shift in the cost norm curve, but since the new cost norm is lower than the annuity-based cost for every year, the new profitability will be lower than the cost of capital. Method B tilts the cost norm curve such that is it almost equal to the cost based on book values, and this yields a higher profitability than for method B, although still slightly lower than the cost of capital. Method C is similar to method A, but here we see both a vertical shift *and* some tilting of the cost norm curve. Profitability is still lower than the cost of capital.

Table 9 shows the magnitude of the calibration effects on ex ante revenue caps for the industry in 2007 and 2008. For these two years, the regulated rates of return

[11] See NVE (2006a).

[12] See Bjørndal et al. (2008).

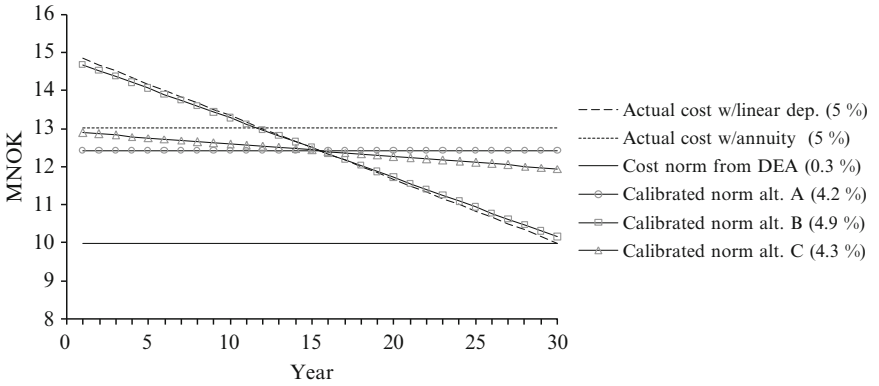


Fig. 11 Cost norms over the life span of a company, with and without calibration

Table 9 Calibration effects for the entire industry in 2007 and 2008

	2007		2008	
	MNOK	"Profitability" (%)	MNOK	"Profitability" (%)
Revenue cap based on DEA eff. scores	12,986	6.54	13,848	6.37
Effect of adjusting efficiency scores (step 2)	599	1.55	786	2.01
Revenue cap after step 2 adjustments	13,585	8.09	14,635	8.38
Compensation parameter (step 3)	328	0.85	371	0.95
Revenue cap before calibration (IR1)	13,913	8.94	15,006	9.33
Calibration effect (step 3)	−328	0.85	−372	−0.95
Final revenue cap (1R2)	13,585	8.09	14,634	8.38

were 8.09% and 8.38%. The calculation of the revenue caps were made in three steps. In Step 1, some minor adjustments, related to the VOLL cost element, were made to the efficiency scores. In Step 2, the efficiency scores were adjusted to bring the industry average up to 100%.[13] In Step 3, the revenue caps were adjusted in order to compensate for the reporting time lags in the regulation model,[12] and then a final calibration was performed in order to bring the industry revenue level (down) to the actual cost level for the industry. We see that the two adjustments in Step 3 cancel each other out at the industry level, although the net effect can be positive or negative

[13] In 2007, calibration method A was used in order to achieve this, and in 2008 method C was used.

for individual companies.[14] The main calibration effect, therefore, was achieved in Step 2, where the industry revenues were increased by 599 MNOK and 786 MNOK, respectively. Note that the corresponding "profitability" effects of 1.55% and 2.01% are somewhat misleading, since the percentages have been calculated with respect to capital reported for year $t-2$; i.e., these are not actual profitability effects. More details on calibration can be found in Bjørndal and Bjørndal (2006b) and Bjørndal et al. (2008b).

4 Concluding Remarks

In this paper, we have given an overview of the benchmarking models used in the Norwegian incentive regulation after 1997. We discuss some specific issues as regards to the development of the DEA models. This concerns model specification issues such as the choice of scale assumptions and input and output variables, but also how the results from the benchmarking models are used, i.e., translated into revenue caps. Due to the considerable effects that model specification and data measurement choices can have on efficiency scores (and thus revenue caps), we argue that in order to specify reasonable DEA models, thorough knowledge about the cost structures and cost drivers of the industry is required. Moreover, we show that in order to achieve the goals of the incentive regulation, some adjustment to the benchmarking results may be necessary before they can be applied in the revenue cap formula.

References

Agrell, P. J., Bogetoft, P., & Tind, J. (2005). DEA and dynamic yardstick competition in scandinavian electricity distribution, *Journal of Productivity Analysis, 23*, 173–201.

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis, *Management Science, 39*(10), 1261–1264.

Banker, R. D., & Natarajan, R. (2004). Statistical tests based on DEA efficiency scores. In W. W. Cooper, L. M. Seiford, & J. Zhu (Eds.), *Handbook on data envelopment analysis* (pp. 299–322). Kluwer.

Banker, R. D., & Chang, H. (2006). The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research*, 175(2), 1311–1320.

Bjørndal, E., Bjørndal, M., & Bjørnenak, T. (2004), *"Effektivitetskrav og kostnadsgruppering"*, Institute for Research in Econ. and Bus. Adm. (SNF), Report 23/04.

Bjørndal, E., & Bjørndal, M. (2006a), *"Nettregulering 2007 – Effektivitetsmåling, gjennomsnittlig effektivitet og aldersparameter"*, Institute for Research in Econ. and Bus. Adm. (SNF), Report 37/06.

[14] In total, the companies were not given compensation for time lags, although, as pointed out by Bjørndal et al. (2008b), they should be. As a consequence, reporting time lags have been removed from the regulation since 2009.

Bjørndal, E., & Bjørndal, M. (2006b). *"Effektivitetsmåling av regional- og distribusjonsnett – fellesmåling, kostnadsvariasjon og kalibrering"*, Institute for Research in Econ. and Bus. Adm. (SNF), Report 38/06.

Bjørndal, E., Bjørndal, M., & Camanho, A. (2008a). *Weight restrictions on geography variables in the DEA benchmarking model for Norwegian electricity distribution companies*. Institute for Research in Econ. and Bus. Adm. (SNF), Report 33/08.

Bjørndal, E., Bjørndal, M., & Johnsen, T. (2008b). *Justeringsparameteren i inntektsrammereguleringen – vurdering av behov for endringer*. Institute for Research in Econ. and Bus. Adm. (SNF), Report 37/08.

Bjørndal, M., & Johnsen, T. (2004). *Nyverdibaserte nettrelaterte kostnader*. Institute for Research in Econ. and Bus. Adm. (SNF), Report 24/04.

Coelli, T. J., Prasada Rao, D. S., O'Donnell, C. J., & Battese, G. E. (2005). *An introduction to efficiency and productivity analysis* (2nd ed.). Springer.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). Data envelopment analysis: a comprehensive text with models, applications, references and dea-solver software (2nd ed.). Springer.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research,* 132, 245–259.

Färe, R., Grosskopf, S., & Zelenyuk, V. (2004). Aggregation bias and its bounds in measuring technical efficiency. *Applied Economics Letters*, 11, 657–660.

von der Fehr, N. -H. M., Hagen, K. P., & Hope, E. (2002). *Nettregulering*. Institute for Research in Econ. and Bus. Adm. (SNF), Report 1/02.

Kittelsen, S. A. C. (1993). *Stepwise DEA; Choosing variables for measuring technical efficiency in Norwegian electricity distribution*. Memorandum 6/1993, Department of Economics, University of Oslo.

Lowry, M. N., & Getachew, L. (2009). Statistical benchmarking in utility regulation: role, standards and methods. *Energy Policy*, 37, 1323–1330.

NVE (1997). *"Retningslinjer for inntektsrammen for overføringstariffene"*.

NVE (2001). *"Den økonomiske reguleringen av nettvirksomheten"*.

NVE (2006a). *"Fastsettelse av kostnadsnorm. Økonomisk regulering av nettselskapene fra 2007"*. Report published 6/6–2006.

NVE (2006b). *"Om fastsettelse av kostnadsnorm for 2007"*. Report published 4/12–2006.

Tauer, L. W. (2001), Input aggregation and computed technical efficiency. *Applied Economics Letters*, 8, 295–297.

On Depreciation and Return on the Asset Base in a Regulated Company Under the Rate-of-Return and LRIC Regulatory Models

L. Peter Jennergren

Abstract This chapter discusses elementary properties of allowed depreciation and return on the asset base for a regulated company under two regulatory models, the traditional rate-of-return model and the more recent LRIC model. Under the former model, any method of computing depreciation and return on the asset base is, in principle, fair to the regulated company and its customers. Under the latter model, only the real annuity method is fair.

1 Introduction and Overview

A regulated company is not allowed to set its prices freely. Rather, there is a regulator that oversees the market and checks whether the prices that are charged are reasonable. In principle, they should be comparable to prices that would be charged in a market characterized by competition. Regulated prices with this property may be referred to as 'market-like'. Regulation means almost by definition that there is no effective competition. The role of the regulator is to create market conditions that resemble effective competition, by imposing market-like prices.

Regulation of an electricity distribution company could, for instance, work as follows. At the end of the year, the company submits information about the number of customers and the amount of energy (kWh) that has been distributed, as well as the sales revenue that it has collected from its customers. The regulator makes a judgment as to the reasonableness of the sales revenue in relation to what has been delivered and orders a refund, if the revenue is found to be too high[1]. In particular,

[1] The example is a stylized description of regulation of electricity distribution in Sweden. It is perhaps somewhat unusual in that the regulation is ex post, that is, the regulator reviews the tariffs *after* they have been applied. According to Jamasb and Pollitt (2007 , p. 5), ex ante regulation is the preferred regime by the majority of regulators. On ex post regulation, see Agrell and Bogetoft (2002). Cf. also Agrell et al. (2005) and EBL Kompetanse AS (2004) on different Scandinavian regulatory regimes in electricity distribution.

L.P. Jennergren
Department of Accounting, Stockholm School of Economics, 11383 Stockholm, Sweden
e-mail: peter.jennergren@hhs.se

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_19,
© Springer-Verlag Berlin Heidelberg 2010

the sales revenue should be broadly in line with the allowed cost of the distribution services, including a reasonable return on the capital that is invested in the distribution network. It is important to stress that here one is concerned with *allowed* cost, that is, cost that would be incurred if the distribution company was subject to effective competition in a well-functioning market. The allowed cost may well be different from what the company itself reports, for example, in its annual financial statements.[2]

The problem for the regulator is apparently to calculate the allowed total cost for the services delivered during a given year. Such a calculation could consist of the following parts: (1) allowed cash cost for operations and maintenance of the network, *plus* (2) allowed depreciation of network equipment, *plus* (3) allowed return on the investment in network equipment.[3] The sum of these three components equals total allowed cost during the year. This chapter is concerned only with the second and third components in such a calculation, that is, allowed depreciation and allowed return. More precisely, the purpose of this chapter is to discuss elementary properties of allowed depreciation and return on the investment under two regulatory models, the rate-of-return model and LRIC (these models will be explained below).

For a regulated company, the *asset base* (sometimes also referred to as *rate base*) is important. The asset base is related to allowed depreciation, since the latter refers precisely to the decrease in the asset base because of wear and tear that the regulated company is allowed to charge for. More precisely, the asset base is that part of the original acquisition price that remains after all of the allowed depreciation so far has been deducted. The asset base is also relevant for the allowed return, since the allowed (absolute) return is the asset base multiplied by some allowed rate of return. The allowed rate of return in this chapter is the nominal weighted average cost of capital (WACC). A method for calculating allowed depreciation and allowed return must, hence, specify the asset base, the allowed depreciation schedule and the WACC.

As for the WACC, in connection with regulation, it is often pre-tax. If so, the allowed absolute return on the asset base is also pre-tax, that is, that return is taxable profit. The reason why the WACC is often pre-tax in connection with regulation seems to be that the revenues that a regulated company collects from its customers and that are subject to the scrutiny of the regulator are pre-tax. In order to earn the allowed return after tax, the company must obtain a higher pre-tax return. This

[2] For instance, in the regulated area of mobile telephony interconnection, around 1999 the Swedish incumbent Telia incurred marketing costs for attracting new subscribers (subsidized handsets). The regulator Post & Telestyrelsen (The Swedish Post and Telecom Agency) did not include this cost item in the allowed cost. (Incumbent is a term sometimes used in the regulation literature to designate a newly privatized, previously state-owned monopoly. Telia merged in 2003 with the Finnish incumbent Sonera to form TeliaSonera.)

[3] These three parts of the allowed total cost can be usefully distinguished in regulatory practice in Sweden that the author is familiar with (electricity distribution and mobile telephony interconnection). The three parts also correspond to the 'building block' approach that is mentioned in Davis (2006, pp. 105–106).

is allowed for by multiplying the asset base by a pre-tax WACC that is equal to the after-tax WACC divided by 'one minus the tax rate'.[4] The WACC is taken to be pre-tax in this chapter, but the distinction between pre-tax and after-tax is not important. Hence, taxes will be disregarded, except for a brief comment in the following section.

From the regulated company's point of view, the allowed return in each year, that is, the asset base multiplied by the WACC, constitutes one cash flow element that it obtains from operations. Another cash flow element is provided by the depreciation that is allowed. Depreciation is usually a non-cash expense. However, in this case it can be regarded as a cash inflow since it enters into the allowed cost that the company is permitted to charge its customers. A third cash flow element, this time an outflow, is given by the initial necessary capital expenditures for new pieces of equipment, that is, for equipment replacements and additions.

To simplify, but with no loss of generality as regards essential insights, depreciation and return on the asset base will, in this chapter, be looked upon from the point of view of *one individual piece of equipment*, rather than from the total stock of equipment (consisting of many individual pieces of different vintages). The asset base will, hence, also refer to one individual piece of equipment.[5]

There are a number of different methods for calculating allowed depreciation and allowed return on the asset base. In connection with regulation, one is particularly interested in methods that are *fair*, meaning the following when one is considering one individual piece of equipment: *The present value of the allowed depreciation and return on the asset base, when discounted at the WACC, should be equal to the initial investment.* If this equality holds, then the regulated company obtains a fair return on its investment, meaning a return that is equal to the WACC. This situation is supposedly similar to what would happen in a market that is characterized by effective competition (see, e.g. Koller et al. 2005, p. 284).

The next section mentions four standard methods for calculating allowed depreciation and return on the asset base. These methods are all fair in the sense mentioned earlier and are often used in connection with the *rate-of-return regulatory model*. This model is based on depreciation and return on the acquisition price that the regulated company actually did pay for each piece of equipment. In other words, under the rate-of-return model the asset base is an accounting book value.[6] The third

[4] In corporate finance, it is not recommended to evaluate, for instance, an investment project by discounting pre-tax cash flows at a pre-tax WACC. In particular, setting the pre-tax WACC to the after-tax WACC divided by 'one minus the tax rate' is no more than an unreliable rule of thumb (as is well known). However, in a regulatory situation, it is not inconsistent to use a pre-tax WACC to calculate a pre-tax allowed return on the asset base, assuming that the regulatory depreciation schedule is equal to the depreciation schedule that is accepted by the tax authority (so that is the assumption made here). See Davis (2006, especially Proposition 1 on p. 112).

[5] Also for simplicity, working capital is disregarded (this could, e.g. be interpreted to mean that working capital assets are exactly balanced by trade credit).

[6] The asset base need not be identical to the accounting book value in the company's own books. If so, the asset base is a different accounting book value, constructed by the regulator. Of course, the regulatory process is simplified if the regulator accepts the company's book value as the asset base.

section discusses allowed depreciation and return on the asset base in connection with the more recent *LRIC* (Long Run Incremental Cost) model for regulation. The fourth and final section comments on relationships between LRIC and the rate-of-return model, in particular, why a regulated company may have a preference for one or the other of the two models. Despite the fact that a very simple analysis framework is used in this chapter, a couple of conclusions can be drawn.[7]

Before moving on to the four standard methods, the notations used in this chapter are given below:

N the economic life of the piece of equipment that is considered (in integer years);

t index for year $1, \ldots, N + 1$;
i the rate of price increases (i.e. inflation), assumed non-negative;
r_r the pre-tax real WACC;
r_n the pre-tax nominal WACC, equal to $((1 + r_r)(1 + i)–1)$;
τ the tax rate.

2 Four Standard Methods for Allowed Depreciation and Return on the Asset Base and Their Use in the Rate-of-Return Regulatory Model

Four standard methods for calculating allowed depreciation and return on the asset base are often discussed in the literature: (a) the nominal linear method, (b) the real linear method, (c) the nominal annuity method, and (d) the real annuity method (see, e.g. Yard 2004). Suppose that a piece of equipment is acquired at the beginning of year 1, and that the economic life is N years. For simplicity, the acquisition price of this piece of equipment is taken to be 1. The distinction between the four methods lies in the asset base at the beginning of year $t = 2, \ldots, N$ (the asset base at the beginning of year 1 is obviously 1 for all the four methods). Consequently, the crucial step in setting up one of these methods is the specification of the asset base. The asset base at the beginning of year t minus the asset base at the beginning of year $t + 1$ determines the depreciation at the end of year t. The return at the end of year t is equal to the nominal rate of return r_n (the nominal WACC) times the asset base at the beginning of year t. The asset base at the beginning of year t and the resulting sum at the end of year t of depreciation and return on the asset base are as follows for each one of the four methods (these formulas are well known from the literature; again see, e.g. Yard 2004).

[7] This chapter's contrasting of rate-of-return and LRIC seems somewhat original. Other papers have usually compared the rate-of-return model to a different alternative, the price cap model (for instance, Beesley and Littlechild 1989; Liston 1993; Joskow 2006; Armstrong and Sappington 2007, pp. 1606–1640).

(a) *The nominal linear method:* the asset base at the beginning of year t is

$$1 - \frac{t-1}{N},$$

and the sum at the end of year t of depreciation and return on the asset base is

$$1 - \frac{t-1}{N} - \left(1 - \frac{t}{N}\right) + r_n \left(1 - \frac{t-1}{N}\right) = \frac{1}{N} + r_n \left(1 - \frac{t-1}{N}\right).$$

(b) *The real linear method:* the asset base at the beginning of year t is

$$\left(1 - \frac{t-1}{N}\right)(1+i)^{t-1}.$$

This value is recognized as the *utilization value* (in Swedish: bruksvärde), equal to the acquisition price at the beginning of year t minus accumulated depreciation on that acquisition price (cf. Johansson and Samuelson 1997, p. 125). The sum at the end of year t of depreciation and return on the asset base is

$$\begin{aligned}
&\left(1 - \frac{t-1}{N}\right)(1+i)^{t-1} - \left(1 - \frac{t}{N}\right)(1+i)^{t} + r_n \left(1 - \frac{t-1}{N}\right)(1+i)^{t-1} \\
&= (1+r_r)\left(1 - \frac{t-1}{N}\right)(1+i)^{t} - \left(1 - \frac{t}{N}\right)(1+i)^{t} \\
&= \left(\frac{1}{N} + r_r \left(1 - \frac{t-1}{N}\right)\right)(1+i)^{t}.
\end{aligned}$$

The last expression in this equation is recognized as the sum of depreciation plus return according to the real linear method.

(c) *The nominal annuity method:* the asset base at the beginning of year t is

$$\frac{(1+r_n)^N - (1+r_n)^{t-1}}{(1+r_n)^N - 1},$$

and the sum at the end of year t of depreciation and return on the asset base is

$$\begin{aligned}
&\frac{(1+r_n)^N - (1+r_n)^{t-1}}{(1+r_n)^N - 1} - \frac{(1+r_n)^N - (1+r_n)^{t}}{(1+r_n)^N - 1} + r_n \frac{(1+r_n)^N - (1+r_n)^{t-1}}{(1+r_n)^N - 1} \\
&= \frac{r_n (1+r_n)^N}{(1+r_n)^N - 1} = \frac{r_n}{1 - (1+r_n)^{-N}}.
\end{aligned}$$

Here, the nominal annuity formula is easily recognized.

(d) *The real annuity method.* This method is sometimes also referred to as the *tilted annuity method* in the regulation literature (see, e.g. NERA Economic Consulting 2006, Appendix A). The asset base at the beginning of year t is

$$\frac{(1+r_r)^N-(1+r_r)^{t-1}}{(1+r_r)^N-1}(1+i)^{t-1},$$

and the sum at the end of year t of depreciation and return on the asset base is

$$\begin{aligned}&\frac{(1+r_r)^N-(1+r_r)^{t-1}}{(1+r_r)^N-1}(1+i)^{t-1}-\frac{(1+r_r)^N-(1+r_r)^t}{(1+r_r)^N-1}(1+i)^t\\&+r_n\frac{(1+r_r)^N-(1+r_r)^{t-1}}{(1+r_r)^N-1}(1+i)^{t-1}=\frac{r_r(1+r_r)^N}{(1+r_r)^N-1}(1+i)^t\\&=\frac{r_r}{1-(1+r_r)^{-N}}(1+i)^t.\end{aligned}$$

Here, one easily recognizes the real annuity formula.

It is easy to see that for all of these four methods, the asset base at the beginning of year 1 is equal to 1, and at the beginning of year $N+1$(i.e. immediately after the end of the economic life) equal to 0. One can show that for $t=2,\ldots,N$ the asset base is strictly larger under the real annuity method than under any one of the other three methods, if $i>0$ and $r_r>0$.[8] It then follows that for $t=1$ the sum of depreciation and return on the asset base is strictly smaller for the real annuity method than for the other three methods. This must hold since for $t=1$ the allowed return is the same under all the four methods (remember that the asset base is 1 for all the four methods at the beginning of year 1). However, the depreciation (being equal to the asset base at the beginning of year 1 minus the asset base at the beginning of year 2) is necessarily smaller under the real annuity method, because the asset base under the real annuity method at the beginning of year 2 is larger than that of the other methods at the beginning of the same year.

It is elementary that under all of the four mentioned methods, the present value, with r_n as the discount rate, at the start of year 1 of the yearly sums of allowed depreciation and return on the asset base is equal to 1, that is, equal to the initial investment. This means that the regulated company's purchases of equipment are fair investments in the sense discussed earlier, that is, zero-NPV transactions. Incidentally, this holds irrespective of whether the discounting is pre-tax or after-tax. For instance, for the nominal linear method it holds that

[8] To show that the asset base under the real annuity method is strictly larger than the asset base under the nominal annuity method is actually somewhat technical. Similarly, to show that the asset base under the real annuity method is strictly larger than the asset base under the real linear method is also somewhat technical. (Proofs can be obtained from the author on request.)

$$\sum_{t=1}^{N}\left(\frac{1}{N}+r_n\left(1-\frac{t-1}{N}\right)\right)\left(\frac{1}{(1+r_n)^t}\right)$$
$$=\sum_{t=1}^{N}\left(\frac{1}{N}+r_n(1-\tau)\left(1-\frac{t-1}{N}\right)\right)\left(\frac{1}{(1+r_n(1-\tau))^t}\right)=1.$$

In the first summation the WACC is apparently pre-tax, and in the second it is after-tax.

More generally, under any one of the four methods, the present value at the start of year t, with r_n as the discount rate, of subsequent payments of depreciation and return on the asset base is equal to the asset base according to the specific method at the start of that year. This means that the asset base is also an economic value that results from a valuation operation (i.e. from the discounting of subsequent payments of depreciation and return). This property actually holds not only for the four standard methods, but also for any other method of calculating depreciation (as long as the remaining values are 1 at the start of year 1, and 0 at the start of year $N+1$). It may be considered as the foundation of the rate-of-return regulatory model where the asset base is an accounting book value.

It should be noted that the rate-of-return model is not explicit about which one of the four standard methods, or indeed any other fair method for calculating depreciation, is used for specifying allowed depreciation and return on the asset base. However, the real annuity method has the following advantage. Under that method, allowed depreciation and return on the asset base is seen to increase from year to year *in line with assumed inflation*. This holds not only during the economic life of one piece of equipment, but also when the company retires that piece of equipment after N years and acquires a new piece of equipment as a replacement (at an acquisition price $(1+i)^N$ times the acquisition price of that piece that is being retired). In other words, the real annuity method is consistent with the regulated company's revenues increasing over time in line with assumed inflation. (This advantage may not be very important if the company has a stock of equipment with an even distribution of vintages.)

The implication is that regulation under the rate-of-return model is fair, since a regulated company is allowed to earn its nominal WACC (equal to r_n) on its investments in equipment. Several authors have commented on this fairness property of the rate-of-return model. For instance, Navarro et al. (1981, p. 403) call it "remarkable". Schmalensee (1989, p. 293) writes "... even though rate-of-return regulation is based on accounting profitability, rate-of-return regulation is in principle fair to both investors and rate-payers *no matter how depreciation is computed* [emphasis in original]. More precisely, if a regulated firm is allowed to earn its actual (nominal) one-period cost of capital on the depreciated original cost of its investments, and if actual earnings equal allowed earnings, then the net present value of all investments is zero for any method of computing depreciation." On a similar note, Salinger (1998, p. 154) writes

> "It is well established but nonetheless remarkable that traditional regulatory practice started with essentially arbitrary (typically straight-line) depreciation schedules and generated prices that had a key feature of cost-based prices: the revenues following from an investment provided the allowed return to the regulated utility. Starting with any depreciation schedule, the company gets a suitable return provided that it is given the depreciation plus the allowed return on the undepreciated portion of the asset in each period."

One may also conclude that (at least in principle) under the rate-of-return model a regulated company's customers have to pay no more than market-like prices for goods and services that they acquire. This follows, since the WACC that is applied is a market required rate of return, or at least estimated from market data.

On the other hand, the rate-of-return model has well-known disadvantages. In particular, it is static and backward-looking. Suppose that owing to technological progress new equipment that can deliver services of a much higher quality becomes available before the end of the originally assumed economic life of an existing piece of equipment. Thus, the original assumption of an economic life of N years that was incorporated into allowed depreciation and return on the asset base may in fact have been incorrect, that is, too long. However, under rate-of-return regulation, which purports to recover historical cash outlays for equipment, there is no incentive for the company to replace an old piece before the N years have passed (cf. Biglaiser and Riordan 2000). In addition to being static and backward-looking, rate-of-return regulation has other weaknesses such as inducing cost padding.[9]

Presumably, owing to these disadvantages, the rate-of-return model has recently lost in popularity. A strong current contender (at least in Sweden) is the LRIC model that is discussed in the following section.[10]

[9] For a survey of drawbacks of rate-of-return regulation with many references to the literature, see Liston (1993). Cf. also Laffont and Tirole (2000, pp. 3, 84–85).

[10] It is not suggested here that LRIC is the only competitor to the rate-of-return model (cf. footnote 7 above). For instance, Biglaiser and Riordan (2000, p. 759) give the following succinct sketch of the development of regulation of telecommunications in the United States: "... rate-of-return regulation has prevailed ... for much of the century. Beginning in the 1980s, price-cap regulation gradually became ascendant. Now ... there are emerging new forms of wholesale price regulation based on long-run marginal cost concepts." The new forms that these authors refer to apparently include LRIC. Also as suggested by this quote, LRIC has been mainly applied in telecommunications. The Swedish Network Performance Assessment Model (NPAM; cf. Sperlingsson 2003; Gammelgård 2004; Jamasb and Pollitt 2007) that is used in the regulation of electricity distribution is mentioned in the following section as an example of LRIC. The reason is that it appears to be somewhat similar from a structural point of view to LRIC models in telecommunications. It is perhaps more accurate to label NPAM as an engineering cost model. Incidentally, a recent Swedish official inquiry recommends a phase-out of the current regulatory model for electricity distribution including NPAM (SOU 2007 :99, in particular pp. 159 and 285). Instead, some variant of rate-of-return regulation is proposed. Cf. also Yard (2008).

3 LRIC and the Real Annuity Method

The LRIC model operates under the assumption that the network is constructed from scratch and all of the network equipment *purchased anew at the start of each year*. The simulated construction of a new, state-of-the-art network is obviously an exercise in engineering design. To price the total network, one uses a price list for all of the components. The assumption that the network is constructed from scratch, with components purchased anew at the start of each year, is only a piece of fiction for the regulatory process. The regulated company is obviously not presumed to act in that fashion.

By simulating a brand new network, LRIC is forward-looking (at least more forward-looking than the rate-of-return model) and may ideally express true marginal costs. This is of course the main attraction. The sum of allowed depreciation and return on the asset base in each year pertains to $t = 1$, that is, *the first year of the economic life* of each piece of equipment. This is a major difference compared to the rate-of-return regulation. Under the latter, depreciation and return on the asset base takes place *over the entire economic life* of a piece of equipment. Under LRIC, *the only relevant year is the first one* since the equipment is thought of as being acquired anew at the beginning of every year.[11]

Now, suppose that the company is actually planning to acquire a piece of equipment (at the price 1) at the start of year 1 and that this piece of equipment is also part of the simulated, state-of-the-art LRIC network. Suppose that the real annuity method is used in the LRIC model for allowed depreciation and return on the asset base in the first year for each successive, fictitious acquisition of the same piece of equipment. Allowed depreciation and return on the asset base in year t is then equal to

$$(1+i)^{t-1} \times \frac{r_r}{1-(1+r_r)^{-N}}(1+i)\,,$$

where the second factor (after the multiplication sign) expresses the first year (i.e. $t = 1$) sum of depreciation and return on the asset base under the real annuity method, and the first factor expresses the acquisition price of the piece of equipment at the start of year t. That is, that acquisition price is 1 when the piece of equipment is purchased at the start of year 1, $(1 + i)$ when purchased at the start of year 2, $(1 + i)^2$ when purchased at the start of year 3, etc. Discounting over the economic life of the piece of equipment,

[11] Needless to say, the LRIC regulatory model is not only concerned with allowed depreciation and return on the asset base. There is also the allowed cash cost for operations and maintenance. Under LRIC, this cost refers to a (fictitious) situation where the company in every year uses a simulated, newly acquired state-of-the-art network. This cost can therefore also be considered as simulated. It could be quite different from the actual operations and maintenance cost that the company is incurring in connection with its currently existing network. As before, however, operations and maintenance cost is not considered in this chapter.

$$\sum_{t=1}^{N}(1+i)^{t-1}\frac{r_r}{1-(1+r_r)^{-N}}(1+i)\left(\frac{1}{(1+r_n)^t}\right)$$

$$=\sum_{t=1}^{N}\frac{r_r}{1-(1+r_r)^{-N}}\left(\frac{1+i}{1+r_n}\right)^t=1.$$

The right-hand side (equal to 1) is the same as the acquisition price for the piece of equipment that the company is actually planning to purchase at the beginning of year 1, so that purchase is a zero-NPV transaction.

In other words, if the LRIC model is used, meaning that each year's sum of allowed depreciation and return on the asset base pertains to a new piece of equipment purchased at the start of that year, and if that sum of allowed depreciation and return on the asset base is computed in accordance with the first year of the real annuity method, then the regulated company obtains a fair rate of return (equal to the nominal WACC) on its investment in equipment.

The property that was just demonstrated, namely LRIC combined with the first year of the real annuity method results in zero-NPV investments in equipment, does not hold if LRIC is combined with the first year of any one of the other three standard methods in the previous section for calculating allowed depreciation and return on the asset base. On the contrary, LRIC combined with the first year of any one of the other three standard methods results in positive-NPV investments in equipment. In other words, allowed depreciation and interest on the asset base are then too high and hence not fair to the regulated company's customers. This is a direct consequence of the property of the other three methods that was noted above, that is, for $t = 1$, the sum of depreciation and return on the asset base is greater for each one of these three methods than for the real annuity method.

There is, hence, an important lesson as regards LRIC: Like in the rate-of-return model, one must specify the method for calculating allowed depreciation and return on the asset base. *LRIC should be combined with the real annuity method.* This lesson has not been missed by the Swedish regulator Energimarknadsinspektionen (The Energy Markets Inspectorate): This authority states that its LRIC model for electricity distribution (called "Nätnyttomodellen"; in English: The network performance assessment model; cf. footnote 10 above) does indeed incorporate the real annuity method for calculating allowed depreciation and return on the asset base.

4 A Comparison of the Two Models

Despite this chapter's modelling simplicity, a couple of conclusions can be drawn in the final comparison of the rate-of-return and LRIC regulatory models. An initial observation is that for a new piece of equipment that has just been acquired, LRIC (combined with the real annuity method) functions like a restricted version of the rate-of-return model. Both models are apparently fair, but LRIC is equivalent to the

rate-of-return model under a specific assumption about method for depreciation and return on the asset base. LRIC, hence, allows for fewer degrees of freedom than the unrestricted rate-of-return model. This would seem to be a slight disadvantage.

In actual regulatory practice, regulated companies sometimes express clear preferences for one of the two models. For instance, as regards mobile telephony interconnection charges, Telia[12] earlier seemed favourably inclined towards LRIC. One reason could be that Telia anticipated a *windfall gain* from the switch from the rate-of-return model (not combined with the real annuity method) to LRIC. It follows from the previous paragraph that the switch to LRIC is equivalent to a switch to the real annuity method for depreciation and return on the asset base while otherwise staying within the rate-of-return model. As is clear from Sect. 2 above, such a switch means that the asset base increases (since the asset base is larger under the real annuity method than under the other standard methods). Also, the asset base is an economic value; so if there is an increase in the economic value, then there is a windfall gain.

On the other hand, regulated electricity distribution companies in Sweden appear to be negatively inclined towards LRIC. One reason could be that LRIC can be more sensitive to specific parameter choices by the regulator than the rate-of-return model. In particular, there is the danger of parameter choices that may result in losses for the companies. For instance, suppose the regulator sets the life N in the real annuity method that is combined with LRIC greater than the actual economic life of the piece of equipment in question. The present value of allowed depreciation and return on the asset base during the actual economic life of that piece of equipment will then be smaller than the original acquisition price. This danger does not arise in the rate-of-return model. That is, even if the choice of life N that is made by the regulator differs from the true economic life, the rate-of-return model is fair. In any case, the present value of depreciation and return on the asset base is equal to the original acquisition price.

Clearly, it is not meaningful to try to provide an exhaustive comparison at this point between the two regulatory models rate-of-return and LRIC. The comparative study of regulatory regimes is a huge topic that has only been touched upon lightly in this chapter. A number of important aspects have been entirely neglected. To cite only two, the modelling has been deterministic, meaning that uncertainty has been disregarded (including uncertainty concerning the regulator's actions). Also, cash cost for operations and maintenance (and interactions between such cost, on the one hand, and depreciation and return on the asset base, on the other) has not been discussed.

There are other, more important qualities of the rate-of-return and LRIC models that cannot be demonstrated within this chapter's very simple analysis framework. In the author's opinion, there is one such quality that should at least be mentioned in closing – model complexity. It is a very serious disadvantage of LRIC models that they are invariably very complex; in fact, usually so complex that they must

[12] On Telia, cf. footnote 2 above.

be constructed by outside consultants. Consequently, they are not fully understood by regulators and regulated companies (or by customers and politicians). In other words, LRIC hinders *transparency*, which is a very desirable property of a regulatory process. In the end, it may be that qualities such as transparency will decide the choice between the rate-of-return model and the LRIC model.

Acknowledgements The author is indebted to the referees and Henrik Andersson (Stockholm School of Economics) for comments. The author is also indebted to Försäkringsbolaget Pensionsgaranti for economic support.

References

Agrell, P., & Bogetoft, P. (2002). *Ex-post regulation*. Preproject 2 – Final report, Sumicsid.

Agrell, P.J., Bogetoft, P., & Tind, J. (2005). DEA and dynamic yardstick competition in Scandinavian electricity distribution. *Journal of Productivity Analysis, 23*(2), 173–201.

Armstrong, M., & Sappington, D. E. M. (2007). Recent developments in the theory of regulation. In M. Armstrong, & R. Porter (eds.), *Handbook of industrial organization* (Vol. 3, pp. 1557–1700). Amsterdam: Elsevier.

Beesley, M. E., & Littlechild, S. C. (1989). The regulation of privatized monopolies in the United Kingdom. *RAND Journal of Economics, 20*(3), 454–472.

Biglaiser, G., & Riordan, M. (2000). Dynamics of price regulation. *Rand Journal of Economics, 31*(4), 744–767.

Davis, K. (2006). Access regime design and required rates of return: pitfalls in adjusting for inflation and tax effects. *Journal of Regulatory Economics, 29*(1), 103–122.

EBL Kompetanse AS (2004). *Network regulation in the Nordic power sector*. Summary, Publikasjon nr 170b-2004, Oslo.

Gammelgård, M. (2004). *The network performance assessment model*. Licentiate's thesis, Department of Electrical Engineering, Royal Institute of Technology, Stockholm.

Jamasb, T., & Pollitt, M. (2007). *Reference models and incentive regulation of electricity distribution networks: an evaluation of Sweden's network performance assessment model (NPAM)*. Working paper CWPE 0747 & EPRG 0718, University of Cambridge, UK.

Johansson, S. -E., & Samuelson, L. A. (1997). *Industriell kalkylering och redovisning (Industrial calculation and accounting, in Swedish)* (9th ed). Stockholm: Norstedts Juridik.

Joskow, P. L. (2006). *Incentive regulation in theory and practice: electricity distribution and transmission networks*. Working paper CWPE 0607 & EPRG 0511, University of Cambridge, UK.

Koller, T., Goedhart, M., & Wessels, D. (2005). *Valuation: measuring and managing the value of companies* (4th ed). Hoboken, New Jersey: Wiley.

Laffont, J. -J., & Tirole, J. (2000). *Competition in Telecommunications*. Cambridge, Massachusetts: MIT.

Liston, C. (1993). Price-cap versus rate-of-return regulation. *Journal of Regulatory Economics, 5*(1), 25–48.

Navarro, P., Petersen, B. C., & Stauffer, T. R. (1981). A critical comparison of utility-type ratemaking methodologies in oil pipeline regulation. *Bell Journal of Economics, 12*(2), 392–412.

NERA Economic Consulting. (2006). *Application of annuity depreciation in the presence of competing technologies II*. Technical report, Melbourne.

Salinger, M. A. (1998). Regulating prices to equal forward-looking costs: cost-based prices or price-based costs? *Journal of Regulatory Economics, 14*(2), 149–163.

Schmalensee, R. (1989). An expository note on depreciation and profitability under rate-of-return regulation. *Journal of Regulatory Economics, 1*(3), 293–298.

SOU 2007:99. (2007). *Förhandsprövning av nättariffer m.m. (Advance approval of net tariffs and other issues, in Swedish)*. Interim report from the Energy net inquiry, Statens offentliga utredningar, Stockholm.

Sperlingsson, M. (2003). *Nätnyttomodellen (The network performance assessment model, in Swedish)*. Master's thesis, School of Economics and Management, Lund University.

Yard, S. (2004). Costing of fixed assets in Swedish municipalities: effects of changing calculation methods. *International Journal of Production Economics, 87*(1), 1–15.

Yard, S. (2008). *Yttrande avseende delbetänkande till Energi nätsutredningen (Statement on the interim report from the Energy net inquiry, in Swedish)*. Report No. 183, School of Economics and Management, Lund University.

Part III
Natural Resources and Logistics

Rescuing the Prey by Harvesting the Predator: Is It Possible?

Leif K. Sandal and Stein I. Steinshamn

Abstract A predator–prey model is used to analyse the case where the prey has been overexploited for a while and therefore is threatened by extinction even along the optimal harvesting path due to depensation in the biological model. It is assumed here, however, that extinction is unacceptable for non-economic reasons. Various sub-optimal rescue operations involving increased harvest of the predator and reduced, or zero, harvest of the prey are therefore considered. The question is how and when it is possible to rescue the prey from extinction by departing from the optimal path. Such sub-optimal policies are not always feasible. If they are feasible, they imply certainly reduced profits and may even produce negative profit. The objective of this chapter is to find the criteria for when a rescue operation is feasible and to explore the dynamics of this situation.

1 Introduction

This chapter gives a short overview of the field called bioeconomic modelling and then introduces the idea of departing from the optimal paths in order to avoid that a species goes extinct. The rationale behind the latter is the assumption that the species at hand has an intrinsic non-monetary value.

The chapter starts with a survey of the development of bioeconomic modelling and how this has been used in order to improve fisheries management. We go briefly through topics that have been given emphasis at various points in time from emphasis on open access in the early days to emphasis on incentive-based management

L.K. Sandal (✉)
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: leif.sandal@nhh.no

S.I. Steinshamn
Institute for Research in Economics and Business Administration, Breiviksveien 40, 5045 Bergen, Norway
e-mail: stein.steinshamn@snf.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_20,
© Springer-Verlag Berlin Heidelberg 2010

and ecosystem-based management these days (Sanchiro and Wilen 2007). This is meant as a benefit to the readers who are not particularly familiar with this field, and is done in the section called Background and motivation. By doing this, we can also categorize and relate the present chapter to the rest of the literature.

Then we go on to our specific contribution in this chapter. We consider a fish-species that is threatened by extinction, even if there is a harvest moratorium in place, due to critical depensation, and how and when it is possible to rescue this species from extinction by increasing the harvest of its main predator. In order to shed light on this situation, we use a bioeconomic predator–prey model. By critical depensation, it is meant that there is a lower critical biomass level below which the species is bound to go extinct. In the general setting, it may very well be the case that some of the optimal paths lead to extinction. Assuming that for various non-economic reasons it is unacceptable to exterminate any species, we seek to determine if it is feasible to rescue the species at stake by leaving the optimal path. In this chapter, we concentrate on the case where the prey is at risk. This implies reduced harvest of the prey and increased harvest of the predator. Such policies are obviously less profitable than staying on the optimal path, and may even imply negative profit. The objective of this chapter is to find the criteria for when it is feasible, and to explore the dynamics of the situation when it is feasible. The main result is that whether saving the prey is feasible or not depends heavily upon the shape of the biological surplus production function. Previously, it has been common to divide this into three categories: compensation, depensation and critical depensation. We introduce a new category in addition by dividing functions characterized by depensation into weak and strong depensation. It is then shown that this distinction is important to determine whether a rescue operation will be successful or not.

2 Background and Motivation

The use of bioeconomic models for the purpose of improving fisheries management dates back to the pioneering works of Gordon (1954) and Scott (1955) among others. Bioeconomic models are dynamic models that combine economics and biology and can, broadly speaking, be divided in three types. Surplus production models (Schaefer 1954, 1957) use the aggregated biomass of the fish stock as the main biological variable and the biological model is basically the relationship between the aggregated biomass and the surplus growth or surplus production of the fish stock. The model can be found both in discrete and continuous time versions. Recruitment models (Ricker 1954) are discrete time models that calculate the next period's harvestable biomass as a function of this period's escapement defined as harvestable biomass minus catch. Year-class models (Beverton and Holt 1957) represent the third type of bioeconomic models, and keep track of each individual year-class from period to period. Year-class models may be both discrete and continuous, although the results are usually presented in discrete time. In this chapter, a continuous time surplus production model will be applied.

The early works focused, in particular, on the problems associated with open access fisheries as this was the typical situation around the world at the time. Free access to a common resource without any kind of property rights results in excessive use of fishing effort and overinvestment in fishing capacity. The reason for this is that the agents have incentives to participate in the fishery as long as it is profitable, and therefore participation will increase until all profitability is exhausted. At this point, the fishery is overexploited, certainly from an economic point of view, and usually also from a biological point of view[1]. Hardin (1968) gave this phenomenon the name "the tragedy of the commons", although the phenomenon had been systematically studied and described already by Jens Warming (1911), and translated into English in Andersen (1983).

Ideas from the literature on fisheries economics have also gradually been adopted in practical management, but at a varying pace and to a varying degree depending upon where in the world you are. Before 1977, open access characterized most waters except a very limited 12-mile zone from the shore. Even inside this zone, usually there was open access for the respective country's citizens. Then, in the late 1970s, more and more countries extended their economic zones significantly to 200-miles. This, however, took place more than 20 years after fisheries scientists such as Scott and Gordon pinpointed the drawbacks of open access. As some sort of property rights, and certainly extended 200-mile zones, are now common, the present chapter takes the viewpoint of a sole owner. More specifically, we think of this as a managing authority that manages the resources for the best of the society.

Other example of ideas from the bioeconomic literature that gradually have found their way into practical fisheries management are individual transferable quotas (ITQs) and so-called incentive-based management (Grafton et al. 2006). ITQs were implemented in the Icelandic herring fisheries in 1976 (Arnason 1993) and after that other Icelandic fisheries followed. New Zealand was another country to adopt ITQs quite early (Lindner 1993). Many other countries like Australia, Canada, and USA have now gradually introduced ITQs in most of their fisheries.

A whole generation of fisheries economists owe their basic knowledge about bioeconomic modelling to Colin Clark's classic book first published in 1976 and revised in 1990 (Clark 1990). The art of bioeconomic modelling has gradually become more and more advanced. One step was to go from steady state analysis, which was common in the 1950s and 1960s, to dynamic optimization as used by Clark and others in the 1970s and onwards. This opened for studying capital-theoretic aspects of fisheries management and the role of discounting; see Clark and Munro (1975) which was followed by a number of papers later, for example, Sandal and Steinshamn (1997, 2001). Another step was to go from one-dimensional single species modelling to more advanced multidimensional multi-species modelling, see Clark (1990) for an introduction to the concept. In the present chapter, we apply a classic two-dimensional predator–prey model.

[1] Economic overexploitation is here defined as harvesting at a rate that is larger than the optimal one. Biological overexploitation is defined as keeping a stock at a lower level than necessary to produce a given biological surplus.

Clark (1990) was also among the first to put emphasis on the importance of depensation (convex part) and critical depensation (critical lower biomass) in the biological growth function as most of the literature on surplus production models up to then had focused on the symmetrical, logistic growth function. In this chapter, we use biological growth functions for both species that can be quite general in the stock biomass and where depensation and critical depensation play a crucial role for the results. Furthermore, we introduce several new concepts. In addition to the distinction between weak and strong depensation, we also introduce the new concept induced critical depensation. This means that a growth function that is characterized by depensation (or even compensation) can be induced to entail critical depensation when a predator–prey interaction term is added.

Multidimensionality can also be introduced by adding other state variables than the biological stock, for example, physical capital. This opens for new kinds of dynamics. In addition to the fundamental dynamics of the natural resource stock comes the dynamics of fishing effort. This can again be divided into capital dynamics and labour dynamics. Smith (1969) was one of the first to look at this phenomenon and propose the rules for entry and exit in fisheries. His work has been followed by a number of papers both on capital dynamics (Charles 1983) and labour dynamics (Charles 1989). The problem of non-malleable capital and irreversible investments is a special topic within the capital dynamics literature, and was first studied in the pioneering work by Clark et al. (1979). This has been followed by papers of Charles and Munro (1985), McKelvey (1985), Boyce (1995) and Sandal et al. (2007). Although Smith's model for entry and exit was based on an open access fishery, much of the later literature on effort dynamics has been on effort dynamics in an optimally regulated fishery. As the model presented here is two-dimensional in biology we will ignore capital dynamics.

Pulse fishing, that is a harvest pattern consisting of regular or irregular pulses instead of an even harvest pattern, was found to be optimal under given circumstances already by Pope (1973) and Hannesson (1975). There are several reasons why a harvest pattern characterized by pulse fishing may perform better than an even harvest pattern, and most of them are related to economies of scale in one way or another. For example, even the simplest kind of non-linearity in the cost function may cause an oscillating fishing pattern to perform better than an even one. The economies of scale can often be related to the production functions used in bioeconomic models, for example, the Schaefer production function.

In year-class (multi-cohort) models, pulse fishing may also be optimal whenever the fishing gear is non-selective. It is a well-known result from year-class models that the highest possible yield is achieved by harvesting each cohort at the age when yield per recruit is maximized. With non-selective gear, however, this is not possible and total yield can be increased by harvesting every N years where N is the age of maximum biomass instead of harvesting a fixed proportion of each year-class every year. This is the phenomenon that is described by Pope (1973) and Hannesson (1975). Whereas the economies of scale argument for pulse fishing is a purely economic one, the non-selective gear argument is basically a combination of a technological and a biological argument.

In this chapter, we present yet another reason why uneven fishing patterns may be optimal based on a truly bioeconomic argument and not based on the economies of scale or multi-cohort modelling. If, for different reasons such as mismanagement, the system is in a state with a poor prey-stock relative to the predator-stock, it may be optimal to increase the harvest of the predator to an extent such that it looks like a bulge in the optimal pattern; especially in stock-space. If a rescue operation, as described in the introduction, becomes necessary for non-economic reasons, this comes in addition to the optimal bulge.

3 The Model

The model is a bioeconomic model with two species interaction where $\{x, y\} \in R_+$ represent prey and predator stock biomass, respectively. The dynamic development of the stocks are governed by the equations [2]:

$$\dot{x}(t) = f(x(t)) - \alpha x(t)y(t) - h_1(t), \quad \dot{y}(t) = g(y(t)) + \beta x(t)y(t) - h_2(t) \quad (1)$$

where f and g denote own growth and α and β are non-negative parameters representing the strength of species interaction. The harvest rates, h_1 and h_2, are here considered as functions of time, t. We will later express the harvest rates in feedback forms, that is, as function of stocks (x, y). The same symbols will be used for the harvest rates when they are expressed in feedback form although the functional forms are different. The functional form will be clear from the context. It is assumed throughout this work that the natural condition $f(0) = g(0) = 0$ is met. Assume further that maximum harvest of the predator is proportional to the stock size by a proportionality factor M. This proportionality factor may be interpreted as the maximum effort available. The harvest of the predator is then constrained by

$$h_2(t) \leq M \cdot y(t).$$

The objective is to maximize the net present value of the two fisheries:

$$NPV = \int_0^\infty e^{-\delta t} \pi(x(t), y(t), h_1(t), h_2(t))dt$$

subject to the constraints above, where π is the combined net revenues and δ is the discount rate. Notice that h_1 and h_2 are control variables and that M is just a constraining parameter. This model is solved numerically using a Hamilton–Jacobi–Bellman approach. In order to look for regions in the predator–prey space where the

[2] Dots denote time derivatives.

optimal policies lead to extinction of the prey, feedback policies must be obtained. As an illustration of this point, we use the case of North-East Arctic cod and capelin in the Barents Sea described in the next section.

3.1 Numerical Example

For illustration, we first present the optimal feedback policies for a predator–prey model of capelin and cod based on results in Agnarsson et al. (2008). This report contains a thorough empirical analysis of these fisheries, and the parameters applied should therefore be quite representative. Let p_i denote the inverse demand function and c_i the cost function for species i. The specification of π is then given by:

$$\pi(y, h_1, h_2) = p_1(h_1)h_1 + p_2(h_2)h_2 - c_1(h_1) - c_2(y, h_2)$$

where more specifically

$$\begin{aligned} p_1(h_1) &= 1, \\ p_2(h_2) &= 12.65 - 0.00839 \cdot h_2, \\ c_1(h_1) &= 0.07 \cdot h_1^{1.4}, \\ c_2(y, h_2) &= 5848.1 \cdot \frac{h_2^{1.1}}{y}. \end{aligned}$$

In this particular case, π is independent of x due to the schooling property of the capelin stock. Schooling behaviour of a stock implies that the density remains constant as the stock size varies. Hence costs are not affected by the stock size as the availability of fish remains more or less constant. The biological submodel described by (1) is specified as follows:

$$\begin{aligned} f(x) &= 0.0018x^2 - 1.19 \cdot 10^{-8} \cdot x^3 \\ g(y) &= 0.00022y^2 - 3.49 \cdot 10^{-11} \cdot y^4 \\ \alpha &= 0.00021 \\ \beta &= 1.82 \cdot 10^{-5}. \end{aligned}$$

The optimal feedback solution for capelin and cod harvest are illustrated in Figs. 1 and 2, respectively. Figure 1 shows the optimal capelin harvest in the two-dimensional cod- and capelin-stock space. For very small cod levels, the optimal harvest plan for capelin is similar to what one would expect in the single-species case, namely a steep rise from the moratorium to the static optimum level. For larger cod stocks, a quite interesting pattern emerges. This pattern consists of considerable harvest at low capelin stocks, then a moratorium over a certain range and then a gradual approach to the static optimum at higher stock levels. It is, in particular, the

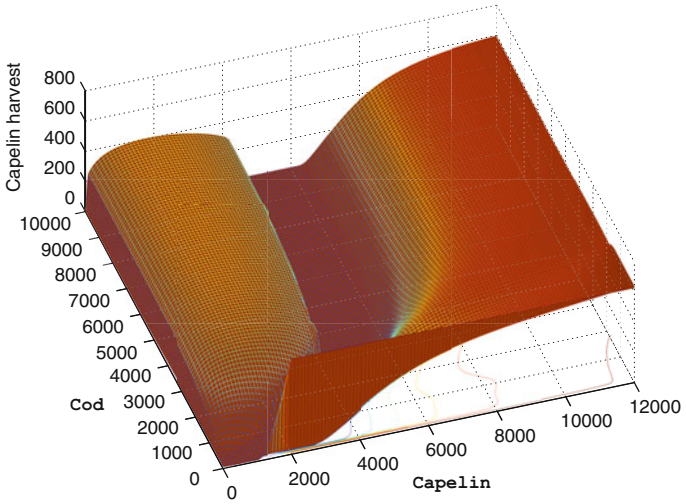


Fig. 1 Optimal feedback solution for capelin harvest in cod–capelin space

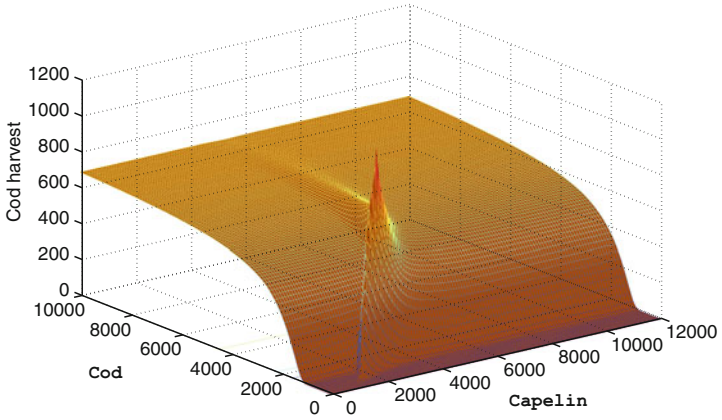


Fig. 2 Optimal feedback solution for cod harvest in cod–capelin space

high harvest at low stock levels that is intriguing because it seems somewhat counterintuitive. The reason for this is that the presence of the cod stock in this model induces critical dispensation along the optimal paths. This is easily seen from Fig. 3 which depicts the dynamic evolution that stems from harvesting according to Figs. 1 and 2. In other words, there is a lower biomass of capelin below which the stock inevitably goes extinct. This lower biomass of capelin is a function of the cod biomass. But, fortunately, it is possible to reduce this area by increasing the cod harvest.

The optimal cod policy in a multi-species perspective is visualized in Fig. 2. Here we can see the optimal harvest of cod for various combinations of the cod and capelin stock. Notice that in most part of this three-dimensional diagram, the

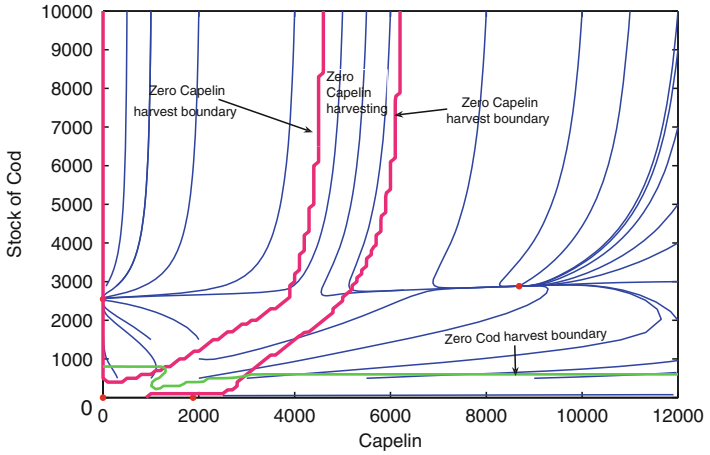


Fig. 3 Flow diagram state space with added optimal zero harvest boundaries

harvest of cod is virtually unaffected by the capelin stock, at least qualitatively. It is similar to what one would expect in the single-species case. However, for certain combinations of cod and capelin stocks, a peak emerges in the diagram indicating that the cod harvest ought to be much higher in this particular area. This bulge represents an optimal rescue operation as opposed to the sub-optimal ones discussed later.

A possible interpretation of the valley in Fig. 1 is twofold: (a) The right side of the valley represents an optimal successful rescue of the capelin. This is seen directly from the directions of the flow in 3. (b) Another aspect of the zero-harvest policy is that it enhances the growth conditions for cod which is the more valuable of the two species. At the left side of the valley this policy cannot prevent the capelin from going extinct, but it will prolong the period when the cod still has another species to prey on.

The next important question is: Is it possible to prevent the capelin from going extinct by departing from these optimal paths? In the next section, we treat this problem in a more general setting.

4 The Rescue Operation

The situation is considered where, in state-space, some of the optimal paths lead to extinction of the prey whereas other optimal paths lead to a steady state where both stocks are positive (positive steady state). Let us call the regions where all optimal paths lead to extinction of one of the species (in this case the prey) the critical regions, and let us call the regions where all optimal paths lead to the positive steady state the non-critical regions. If the fisheries had been optimally managed from the

beginning, starting with pristine stocks, one would have been at the positive steady state. However, due to mismanagement, environmental disasters or other events, it is assumed here that the present state is in the critical region.

It is further assumed that extinction of any of the species is unacceptable for reasons other than pure economic ones. The concern then is to determine under what conditions extinction of the prey is unavoidable and under what conditions the prey may be rescued from extinction by deviating from the optimal path. By a rescue operation is meant deviation from the optimal harvest pattern in the form of not lower than optimal harvest of the predator and not larger than optimal harvest of the prey. A rescue operation is infeasible if extinction is unavoidable even with the extreme policies of maximum harvest of the predator and zero harvest of the prey. This assumes that there is a lower critical biomass level for the prey that depends on the stock level of the predator. Such a critical biomass level is often referred to as critical depensation (Clark 1990), and in this case it can be said that the critical depensation is induced by the predator. The reason for this is that the interaction between the species may cause a change from non-critical single species production to critical effective production. If extinction is unavoidable, optimal management merely consists of optimally mining out the remaining part of the prey resource, as the stock then is beyond rescue anyway.

The critical regions are characterized by $\lim_{t\to\infty} x(t) = 0$ for all optimal paths starting in this region. The question then is: is it possible to bring the stocks out of this region using feasible but sub-optimal harvesting strategies? In order to answer these questions, we will apply state-space analysis.

5 State-Space Analysis

Figure 4 depicts structures in state space based on a generalized version of North-East Arctic cod and capelin in the Barents Sea given by $f(x) = ax^2 - bx^3$ and $g(y) = Ay^2 - By^4$ where $a,\ b,\ A$ and B are parameters. It is assumed that the harvest of the prey is zero, $h_1 = 0$, whereas the harvest of the predator is $h_2 = M \cdot y$. The left panel shows four situations with different effort capacities. The right panel indicates the flow directions in the limiting case. It is important to note that the $\dot{y} = 0$ curve shifts to the right as M increases. This yields three principal situations: (a) the case where the $\dot{y} = 0$ curve intersects with the vertical axis, (b) the case where the vertical axis is tangent to the $\dot{y} = 0$ curve (limiting case) and (c) the case where there is no common point between the two.

The case with intersection with the y-axis is important because it implies that the prey will go extinct even with an extreme rescue operation ($h_1 = 0,\ h_2 = My$). The reason for this is the existence of a stable equilibrium with zero prey. In this case, rescuing the prey from extinction is simply not a feasible option if the predator is abundant.

The next principal case is when the vertical axis is tangent to the $\dot{y} = 0$ curve. This is the limit case with only one equilibrium point. The case with no stable

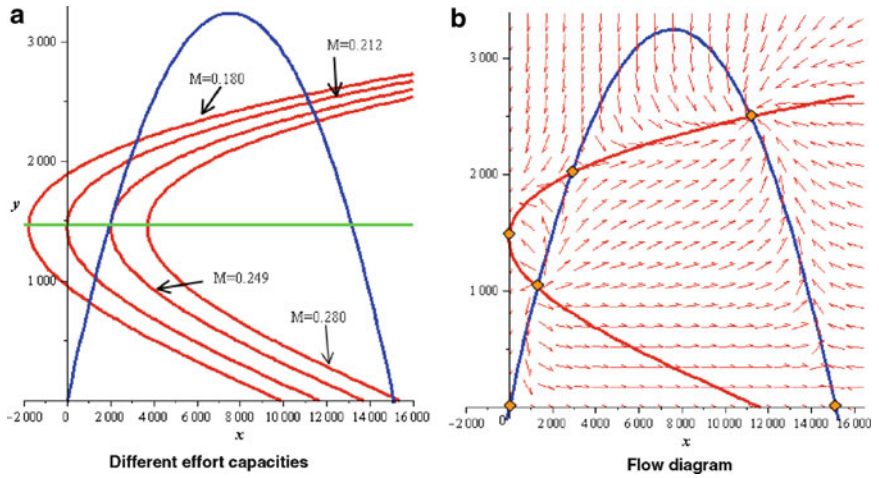


Fig. 4 Phase space for maximum effort on the predator and moratorium on the prey. The A-shaped curve is $\dot{y} = 0$ and the other curve is $\dot{x} = 0$ for non-zero stocks

equilibrium points on the predator axis implies that it is always feasible to avoid that the prey becomes extinct[3]. This reasoning is based purely on biological and technical analysis. A rescue operation may still be extremely expensive from an economic point of view although it is feasible.

This argument is derived directly from the biological submodel (1) by inserting $h_1 = 0$ and $h_2 = M \cdot y$. The basic curves in the phase space are then given by

$$\begin{aligned} \dot{x} = 0 \Rightarrow \left\{ \alpha y = \frac{f(x)}{x} \equiv F(x) \text{ or } x = 0 \right\} \text{ and} \\ \dot{y} = 0 \Rightarrow \left\{ \beta x = -\frac{g(y)}{y} + M \equiv M - G(y) \text{ or } y = 0 \right\}. \end{aligned} \tag{2}$$

It is straightforward to see that the important cases can be sorted out by removal of stable equilibrium points on the vertical axis in the feasible region for the predator stock. The functions F and G represent the mean surplus growth functions for the prey and the predator, respectively.

The logistic case is illustrated in Fig. 5. Logistic versions of f and g can be defined by

$$f(x) = rx\left(1 - \frac{x}{k}\right) \quad \text{and} \quad g(y) = Ry\left(1 - \frac{y}{K}\right)$$

where r, R, k and K are parameters. These parameters, however, lose their usual interpretation as intrinsic growth rates and carrying capacities (see Clark 1990) due to the species interaction. The main results in the logistic case can be summarized in a proposition.

[3] The technical proof showing that the origin is avoided is given in appendix 2.

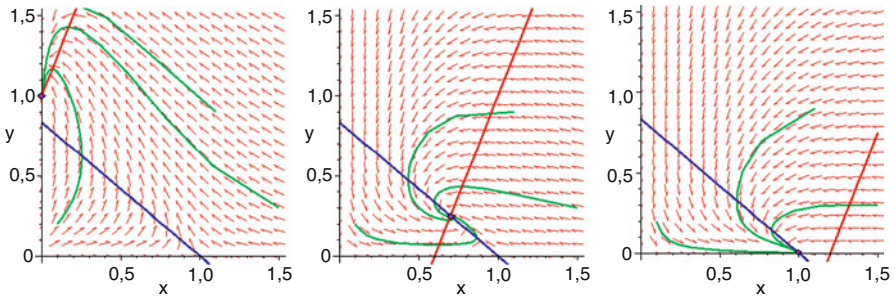


Fig. 5 Principal cases with extreme rescue operation in the logistic growth case

Proposition 1. *(The logistic case) Let f and g be the logistic functions and let the limiting effort that can be applied to the predator be defined by $M^* = R\left(1 - \frac{r}{\alpha K}\right)$. If the prey stock at a certain point in time is severely depleted but not extinct, then the following two mutually excluding situations can arise:*

A: Both stocks can be sustained if the maximum fishing effort that can be applied to the predator is larger than the limiting effort; $M > M^*$. This is always the case if the system allows for a natural coexistence, given other combinations of prey and predator stocks.

B: The prey stock is bound to go extinct if the maximum fishing effort that can be applied to the predator is smaller than or equal to the limiting effort; $M \leq M^*$.

The proof of Proposition 1 is given in Appendix 1. In case A, the maximum possible harvest rate is sufficiently high to reduce the predator stock and allow the prey stock to start growing. After a certain period with maximum effort one should revert to the optimal harvesting scheme, and this should definitely be done before the predator is put at risk. It should, in fact, be done as soon as the prey stock is outside the critical region as defined earlier.

In case B, even the maximum effort available to be applied to the predator is not sufficient for the rescue operation to be successful. This case can also be regarded as a case of critical depensation. As the prey stock in this case is beyond saving, one might as well mine it out in an optimal manner. The optimal mining procedure is given by the optimal paths in the critical region.

The case in the left plot in Fig. 5 represents a special situation with no coexistence without harvesting. This can be due to some kind of environmental regime shift that has changed the biological conditions from a natural sustainable coexistence to a natural removal of the prey. This is brought about by a change in the biological key parameters such that $r/\alpha - K$ changes sign from positive to negative and hence M^* changes sign in the opposite direction.

The phase plane for the logistic case can be divided into three mutually excluding cases as illustrated in Fig. 5. The condition in the left plot illustrates the case when $M \leq M^*$, the middle case illustrates the situation when M is somewhat larger than M^* and right case illustrates a situation when M is much larger than M^*. The

case in the right plot clearly illustrates the importance of reverting to an optimal harvesting regime as soon as the prey stock is brought within safe biological limits as a continuation of the extreme rescue operation when it is no longer required will eventually eliminate the predator stock. From the middle case, it is seen that the predator will never be eliminated in this case even if the extreme rescue policy is continued forever, but it will, of course, be highly unprofitable to do so from an economic point of view. Notice that the origin is a source point on the left plot and a saddle point in the other plots.

Definition 20.1. (Natural growth function). A natural growth function for a biological stock $f : R_+ \to R$ is a function with the following properties:

(a) There exists a carrying capacity $\kappa > 0$ such that $f(\kappa) = 0$ and where f changes sign from positive to negative.
(b) The growth function f is single-peaked.
(c) For small arguments, the growth function is of type $f(x) = x^{\nu}(c_0 + c_1 x + ...)$ with $c_0 \neq 0$ and $\nu > 0$.

Notice that the definition above allows for certain irregularities close to the origin if $\nu > 0$. The definition above includes a large group of possible growth functions and is therefore not very restrictive. The pure logistic growth function applied in Proposition 1 is a simple and widely used natural surplus growth function that is a special case of $f(x) = x^{\nu}(c_0 + c_1 x)$ with $\nu = 1$, $c_0 > 0$ and $c_1 < 0$. The growth functions for North-East Arctic cod and capelin in the Barents Sea in the numerical example are also natural growth functions.

There is one important property of the growth function that may introduce a critical lower biomass. Growth functions that are convex for small stock levels are often said to be characterized by depensation. Critical depensation is used to describe growth functions that are negative for small stock levels (see Clark 1990). Critical depensation implies that there exists a minimum critical biomass below which the stock will inevitably go extinct. We will expand the term depensation to weak, strong and critical according to the following definition.

Definition 20.2. (Weak, strong and critical depensation). A natural surplus growth function, f, is characterized by depensation if it is convex for small stock levels. The depensation is categorized by its slope at the origin as weak if $f'(0) > 0$, strong if $f'(0) = 0$. We use the term critical depensation if f is negative in the vicinity of $x = 0$ whether it is convex or not[4].

Figure 6 illustrates the four principal categories of natural growth functions: With compensation (concave) and the three cases with depensation described above. The mean growth curve, $F(x) = \frac{f(x)}{x}$, is the key determinant for the sign of the growth rate. This can be seen directly from the relations in (2). Now Proposition 1 can be generalized and it is then possible to deal with many more combinations of natural surplus growth functions.

[4] In the case of $0 < \nu < 1$ we define $f'(0) = F(0) = \infty$ if $c_0 > 0$ and $f'(0) = F(0) = -\infty$ if $c_0 < 0$.

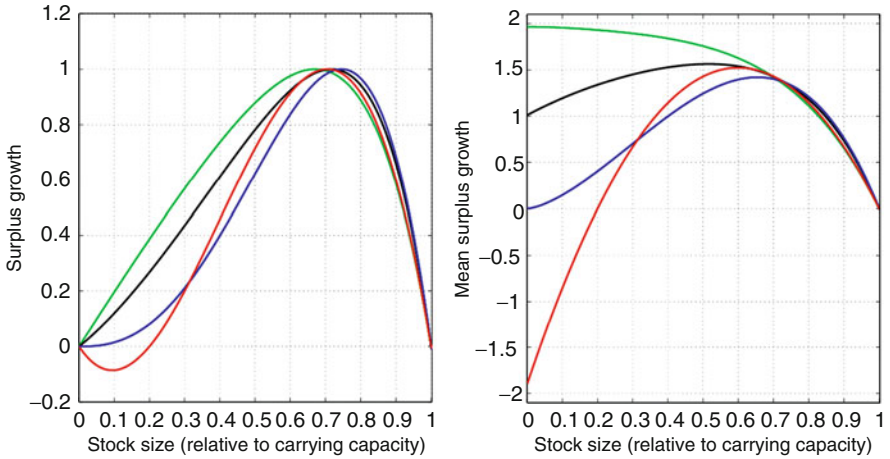


Fig. 6 Plots of a natural surplus growth function $f(x)$ and mean surplus growth function $F(x) = \frac{f(x)}{x}$ for the four cases: compensatory growth, weak, strong and critical depensatory growth. They are depicted in the same order close to the origin. The upper one represents the compensatory growth

Proposition 2. *(Sustainable prey options) Assume that the natural surplus growth functions for the prey and the predator allow for a natural coexistence of both species in the absence of harvesting and assume that the prey is heavily depleted. Then there are four principal cases referring to the growth function of the prey:*

(a) Compensation: There is always a feasible policy that ensures sustainability of the prey
(b) Weak depensation: There is always a feasible policy that ensures sustainability of the prey if $M > M^* = G(\alpha^{-1}F(0))$.
(c) Strong depensation: There is always a feasible policy that ensures sustainability of the prey if $M > M^* = \max G(y)$ *and the predator has depensatory growth in the vicinty of* $y = 0$*, that is,* $\nu \geq 1$*. On the other hand, if* $0 < \nu < 1$ *(compensatory growth) it induces a critical depensation in the effective growth of the prey.*
(d) Critical depensation: It is impossible to ensure sustainability of the prey after it is below the minimum critical biomass.

In connection with Proposition 2, it is interesting to note that by skipping the assumption about coexistence in the absence of harvesting, the case with compensation and the case with weak depensation can be merged, and the condition $M > M^* = G(\alpha^{-1}F(0))$ holds for both cases.[5] It is easily seen that this criterion coincides with the criterion given in Proposition 1 for the logistic case.

[5] This situation may arise if, for example, there is a regime shift affecting the growth conditions such that the prey will go extinct in the absence of harvesting of the predator.

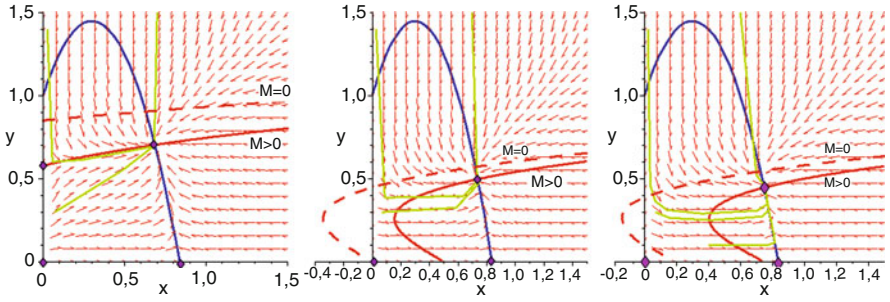


Fig. 7 The principal cases for $F(0) > 0$. The prey can always be sustained. The dotted curve is Γ_y with no harvesting. Predator growth is with (**a**) compensation or weak depensation, (**b**) strong depensation and (**c**) critical depensation

How Proposition 2 is derived is outlined in the following. Our starting point is (2). We define Γ_x as the curve where $\dot{x}$ changes sign and Γ_y the curve where $\dot{y}$ changes sign. The assumption about natural coexistence implies that the two curves intersect in the interior of the first quadrant and form a stable steady state equilibrium.

The simplest cases are sorted out first. They arise when the natural growth function for the prey is compensatory or weakly depensatory and are all of the type $F(0) = f'(0) > 0$. In Fig. 7, the zero growth predator curve Γ_y is the curve going up towards the right. The dotted curve represents Γ_y in the case with no harvest ($M = 0$). The phase plots demonstrate clearly that the prey can be sustained when $F(0) > 0$. This can be achieved without necessarily applying maximum fishing effort to the predator. In all the three cases, the predator stock may be threatened if maximum fishing effort is applied too long such that it makes the intersection between Γ_y and the x-axis to the right of Γ_x, that is, if M is large enough. It is the part of the phase space close to the y-axis and typically below the natural equilibrium that is of interest. Further away from the axis, there is no point in deviating from the optimal policy.

The case with critical depensation for the prey is straightforward. Once the prey stock is below the single species viable stock it is bound to go extinct no matter what the size of the predator stock is. This follows directly from $\dot{x} \leq f(x) < 0$ if $0 < x < x_{crit}$.

The case with strong depensation for the prey is in principle given in Fig. 8. If Γ_y crosses the vertical y-axis as depicted in (a) we have a stable equilibrium at the crossing. It attracts all paths starting with a small prey stock for any size of the predator stock. If $M > G(0)$, this attracting equilibrium is removed. In the fold-back case depicted in (b), one has a stable equilibrium with no prey at a relatively high predator stock. With sufficiently high M, the situation becomes like the one depicted in (c). This case is further discussed in Appendix 2. In the numerical case given for the cod and capelin in the Barents Sea we get $M^* = \max_{y \geq 0} G(y) = 0.212$. Both growth functions have strong depensation. It implies that a maximum fishing effort

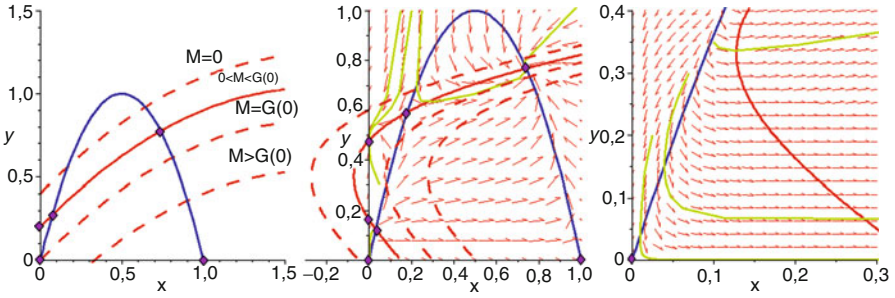


Fig. 8 The principal cases for different M-values. Dotted curves are alternative M-values and flow arrows are for the case with the marked equilibrium points

or fishing mortality (on cod) greater than 21.2% will prevent capelin from going extinct. The result is interesting keeping in mind that it is economically optimal to remove all capelin if we start inside a relative large critical region depicted in Fig. 3.

6 Summary

In this chapter, we have analyzed under what circumstances it is possible to rescue a prey from extinction by increasing the harvest on its main predator even beyond the levels that are optimal and beyond the levels that are profitable at all, assuming that extinction is unacceptable for various non-economic reasons. The objective has been to investigate under what conditions such a rescue operation is feasible or not. We have also analysed the case where a peak in the harvesting pattern for the predator is part of an optimal policy as in the numerical example.

Our main theoretical contribution has been to define a maximum effort level based purely on biological information such that if the maximum effort applied to the predator is higher than this, then the prey stock can be rescued from extinction whereas if the effort applied is less than this, the prey stock is bound to go extinct. The main result is that whether it is possible to save the prey from extinction or not, when it is already severely overexploited, depends on the functional form of the biological growth function, which again is an empirical question. With compensation and weak depensation it is relatively easy to save the prey, with strong compensation it is possible but more demanding and with critical depensation it is impossible after the stock falls below the critical level. The distinction between weak and strong depensation is an original contribution in this chapter and it turns out to be very useful.

In addition we have illustrated the theoretical model with a numerical example based on data for cod and capelin in the Barents Sea showing that for capelin there

is a lower stock level below which there is no point in trying to rescue the stock anymore due to the critical depensation induced by the cod stock through an optimal management regime. For higher stock levels a harvest moratorium must be issued and for even higher stock levels the optimal harvest pattern will be fairly similar to the single-species case although adjusted for the species interaction. This situation is purely economically rooted since the capelin can be rescued as long as the maximum potential fishing effort exceeds 21.2%.

The optimal cod harvest is quite similar to the single species case in most of the combinations except in the area where capelin is subject to moratorium. In this area, optimal cod harvest is characterized by a peak representing a harvest that is even higher than the static optimum. The purpose of this peak is to rescue the capelin by reducing the predation pressure.

7 Appendix: Proof of Proposition 1

Let us rescale the stock variables by $x = ur/\beta$ and $y = vKr/R$ and time by $\tau = rt$ to obtain

$$u' = u(1 - au - bv) \text{ and } v' = v(u - v - q)$$
$$\text{with } r = ak\beta,\ bR = \alpha K \text{ and } qr = M - R.$$

The equilibrium points for this system are $(0,0)$, $A = (a^{-1}, 0)$, $B = (0, -q)$ and $C = (u^*, v^*)$ with $u^* = (1 + bq)/(a + b)$ and $v^* = (1 - aq/(a + b)$. Only the first quadrant is of interest. Two of them do not move around when M is changing, namely the origin and A. Only q is changing when M is changing. The maximum number of equilibrium points is four. It occurs in the intermediate case in Fig. 5 when the two straight lines cross creating a coexistence of both stocks in an equilibrium. We call the stable point S. When q is changing, the only thing that happens is that the upward sloping line is shifted left or right according to the sign of the change in q. S is the only stable equilibrium point. When q is increasing, $S = C$ moves along the downward sloping line until it collapses into the unstable equilibrium A and stays there and dominates A to form a stable point $S = A$. If q is decreasing, the stable point $S = C$ moves along the downward sloping line until it hits the v-axis when $S = C$ and B collide and form a stable point $S = B$ that continues to move up on the axis. The stabilities of the equilibrium point are easily seen from the fact that, for positive stocks, the following hold true: The flow is downwards left of the upward sloping line and upwards to the right of it. The flow is leftward above the downward sloping line and rightward below. B is present if $q < 0$ and C is present if $1 > \max(aq, -bq)$. The graphical arguments above are easily backed up by linear stability analysis. The situation $M = M^*$ occurs when B and C collide. It implies $u^* = 0$ and $-q = v^*$ which is equivalent with $1 + bq = 0$ resulting in the given expression for M^*.

8 Appendix: Analysis of Steady States With No Prey

We look at the model $\dot{x} = x \cdot (F(x) - \alpha y)$ and $\dot{y} = y \cdot (G(y) + \beta x - M)$. Linearization around a steady state (x^*, y^*) can be done using the Jacobi-matrix given by

$$J(x,y) = \begin{bmatrix} \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial y} \\ \frac{\partial y'}{\partial x} & \frac{\partial y'}{\partial x} \end{bmatrix} = \begin{bmatrix} F(x) + xF'(x) - \alpha y & -\alpha x \\ \beta y & G(y) + yG'(y) + \beta x - M \end{bmatrix}.$$

As we focus on extinction of the prey, we are primarily interested in steady states on the y-axis. At the origin we have

$$J(0,0) = \begin{bmatrix} F(0) & 0 \\ 0 & G(0) - M \end{bmatrix} \text{ with eigenvalues } \lambda_1 = F(0) \text{ and } \lambda_2 = G(0) - M.$$

This shows that as long as $F(0) > 0$ we have saddle-point when $G(0) < M$ and a source when $G(0) > M$. Both cases represent stability in the sense that they ensure the presence of the prey. The approaching separatrix in the saddle-point case is the y-axis. The stock will not disappear into the equilibrium point unless we are on the axis already. This is illustrated in Figs. 5 and 7. The case $F(0) < 0$ is trivial as it corresponds to critical depensation for the prey. If $0 < \nu < 1$ we have $F(0) = \infty$ and it implies to leading order that $\dot{x}/x = a_0 x^{\nu-1} - \alpha y = 0$ if $a_0 = \alpha y x^{1-\nu}$. Below this curve, we have $\dot{x} > 0$ implying that the prey stock increases in the vicinty of the origin.

Next we look at other possible steady states on the y-axis. One such point is defined by $x = 0$ and $G(y) - M = 0$, see for example, Fig. 7. By inserting $(0, y^*)$ where y^* is a positive solution of $M = G(y^*)$ we get

$$J(0,y^*) = \begin{bmatrix} F(0) - \alpha y^* & 0 \\ \beta y^* & y^* G'(y^*) \end{bmatrix}$$

with eigenvalues $\lambda_1 = F(0) - \alpha y^*$ and $\lambda_2 = y^* G'(y^*)$.

Notice that the intersection between the $\dot{x} = 0$ curve and the y-axis is given by $F(0) - \alpha y = 0$. Let this be called $Y = F(0)/\alpha$. The intersection between the $\dot{y} = 0$ curve and the y-axis is given by y^*. If, like in Fig. 7, $\dot{x} = 0$ intersects higher than $\dot{y} = 0$, we have $y^* > Y$ implying that $\lambda_1 < 0$ and vice versa. In the case $\lambda_1 > 0$ and $\lambda_2 < 0$, we have a saddle point, and if both eigenvalues are positive we have a source. The stocks will not go extinct in any of these cases. The direction for λ_1 is along the y-axis and for λ_2 it is inside the positive quadrant.

The remaining problem, not shown in Fig. 7, is when $\lambda_1 < 0$. If also $\lambda_2 < 0$, we have a stable steady state on the y-axis implying that the prey will go extinct. As it is assumed that $y^* > 0$, it means $G'(y^*) < 0$. This is equivalent to stating that $y^*(M)$ decreases with M. In the opposite case, there is no problem. This leaves us with a knife-edge case, namely the capacity $M = M^*$ that yields $Y = y^*$. This is

equivalent to $M^* = G(y^*) = G(\alpha^{-1}F(0))$ and in accordance with the results in Proposition 1 regarding the logistic case, namely $M^* = G(y^*) = R(1-y^*/K) = R(1-\alpha^{-1}F(0)/K) = R(1-r/(\alpha K))$. It is also in accordance with the weak depensation part of Proposition 2. For this to be interesting it must entail $y^*(0) > Y$, or, in other words, that $G(y^*) = 0$ yields a solution larger than $F(0)/\alpha > 0$. But the only point $G = 0$ is at the natural carrying capacity (and at a possible lower critical biomass that we do not consider here). Hence, this case is only interesting when $\alpha K > F(0)$.

The case $F(0) < 0$ is trivial as it corresponds to critical depensation for the prey. The case with strong depensation is by defintion $F(0) = 0$. In this case, however, one of the eigenvalues is zero, and the Jacobi-matrix is of no use. We have a non-hyperbolic equilibrium. Due to the approximation (linearization) this is a knife-edge case, and we do not know which side we would be without the simplification.

The case with strong depensation for the prey is in principle given in Fig. 8. If Γ_y crosses the vertical y-axis as depicted in (a) we have a stable equilibrium at the crossing. It attracts all paths starting with a small prey stock for any size of the predator stock. If $M > G(0)$ this attracting equilibrium is removed. In the fold-back case depicted in (b), one has a stable equilibrium with no prey at a relatively high predator stock. With sufficiently high M the situation becomes like the one depicted in (c). The question that needs to be answered is as follows: Will some paths starting in the vicinity of the origin with both stocks present, terminate at the origin or will all such paths leave the neighborhood? To answer this question we must analyze the situation where the predator stock is monotonically decreasing and check whether the path crosses Γ_x into the region where the prey is increasing. We expect this to be true if we are sufficiently far from Γ_y implying that the path is steeper than Γ_x and hence is crossing it. Let $0 < x < \epsilon$ and $0 < y < \epsilon$ for sufficiently small ϵ define a neighborhood U_ϵ of the origin. In this neighborhood, we have $f(x) = x^{\eta+1}(a_0 + a_1 x + ...)$ with $a_0 > 0$ and $\eta > 0$. The latter holds since we are only looking at the case with strong depensation. The dynamic relations $\frac{\dot{x}}{x} = F(x) - \alpha y$ and $\frac{\dot{y}}{y} = G(y) + \beta x - M$ can than be written as $\frac{\dot{x}}{x} = a_0 x^\eta - \alpha y + o(\epsilon)$ with $a_0 > 0$ and $\frac{\dot{y}}{y} = y^{\nu-1}(c_0 + c_1 y + \ldots) + \beta x - M$. We sort out the cases $0 < \nu < 1$ and $\nu \geqq 1$.

Case $\nu \geqq 1$, that is, $G(0) = 0$ or $G(0) = c_0 > 0$.

In this case, we have $\frac{\dot{y}}{y} = G(y) - G(0) + \beta x - q$ where $q = M - G(0) > 0$ in the relevant case where y is decreasing. Keeping track of the leading order the contributions implies $\frac{\dot{y}}{y} = -q + o(\epsilon)$. We form the state path equation $\frac{dx}{dy} = \frac{-ax^{\eta+1}}{y} + bx$ with $(a, b) = (\frac{a_0}{q}, \frac{\alpha}{q})$. The path starting at $(x_0, y_0) \in U_\epsilon$ is given by $x^{-\eta} \cdot e^{b\eta y} = x_0^{-\eta} \cdot e^{b\eta y_0} - a\eta \int_{b\eta y}^{b\eta y_0} s^{-1} e^s ds$. In our case, y is decreasing and the integral term goes to $+\infty$ as $y \to 0^+$. Hence there exists a $0 < \widetilde{y} < y_0$ such that the right hand side becomes zero implying that $x \to +\infty$ as $y \to \widetilde{y}^+$. We can conclude that the path leaves the $U\epsilon$-neighborhood of the origin as depicted in Fig. 7. If M is not large enough to remove the upper stable equilibrium on the y-axis one may end up with no prey if the initial predator stock is large. The removal occurs when M is above $M^* = \max(G(y))$.

Case $\nu \in (0,1)$, that is, $G(0) = +\infty$.

The leading order contribution in the predator equation is then $\frac{\dot{y}}{y} = y^{\nu-1}c_0 - M + o(\epsilon)$ with the leading order solution as $y^{1-\nu} = c_0/M + (y_0^{1-\nu} - c_0/M)e^{-M(1-\nu)t}$ and hence $y \to y_* = (c_0/M)^{1/(1-\nu)}$ for all feasible starting values $y_0 \geq y_*$. From the relation $\frac{\dot{x}}{x} = a_0 x^\eta - \alpha y$ we immediately see that $\dot{x} < 0$ for $x < (ay_*/a_0)^{1/\eta}$. and the prey goes extinct. If $y_0 < y_*$ we have $\dot{x} < 0$ for $x < (ay_0/a_0)^{1/\eta}$. Hence the prey goes extinct if $x < x_* = \min\{(ay_*/a_0)^{1/\eta}, (ay_0/a_0)^{1/\eta}\}$.

References

Agnarsson, S., Arnason, R., Johannisdottir, K., Ravn-Johnsen, L., Sandal, L. K., Steinshamn, S. I., & Vestergaard, N. (2008). Multispecies and stochastic issues: comparative evaluation of the fisheries policies in Denmark, Iceland and Norway, TemaNord 2008:540, Nordic Council of Ministers, Copenhagen.

Andersen, P. (1983). On rent of fishing grounds: a translation of Jens Warming's 1911 article with an introduction. *History of Political Economy, 15*, 391–396.

Arnason, R. (1993). The Icelandic individual transferable quota system: a descriptive account. *Marine Resource Economics, 8*(3), 201–218.

Beverton, R. J. H., & Holt, S. J. (1957). *On the dynamics of exploited fish populations, fisheries investigation series 2(19)*. London: Ministry of Agriculture, Fisheries and Food.

Boyce, J. R. (1995). Optimal capital accumulation in a fishery: a nonlinear irreversible investment model. *Journal of Environmental Economics and Management, 28*, 324–339.

Charles, A. T. (1983). Optimal fisheries investment under uncertainty. *Canadian Journal of Fisheries and Aquatic Sciences, 40*, 2080–2091.

Charles, A. T. (1989). Bio-socio-economic fisheries models: labour dynamics and multiobjective management. *Canadian Journal of Fisheries and Aquatic Sciences, 46*, 1313–1322.

Charles, A. T., & Munro, G. R. (1985). Irreversible investment and optimal fisheries management: a stochastic analysis. *Marine Resource Economics, 1*, 247–264.

Clark, C. W. (1990). *Mathematical bioeconomics: the optimal management of renewable resources*. New York: Wiley.

Clark, C. W., Clarke, F. H., & Munro, G. R. (1979). The optimal exploitation of renewable resource stocks: problems of irreversible investment. *Econometrica, 47*, 25–47.

Clark, C. W., & Munro, G. R. (1975). The economics of fishing and modern capital theory: a simplified approach. *Journal of Environmental Economics and Management, 2*, 92–106.

Gordon, H. S. (1954). The economic theory of a common property resource: the fishery. *Journal of Political Economy, 62*, 124–142.

Grafton, R. Q., Arnason, R., Bjorndal, T., Campbell, D., Campbell, H. F., Clark, C. W., Connor, R., Dupont, D. P., Hannesson, R., Hilborn, R., Kirkley, J. E., Kompas, T., Lane, D. E., Munro, G. R., Pascoe, S., Squires, D., Steinshamn, S. I., Turris, B. R., & Weninger, Q. (2006). Incentive-based approaches to sustainable fisheries. *Canadian Journal of Fisheries and Aquatic Sciences, 63*(3), 699–710.

Hannesson, R. (1975). Fishery dynamics: a North-Atlantic cod fishery. *Canadian Journal of Economics, 8*, 151–173.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1247.

Lindner, R. K., Campbell, H. F., & Bevin, G. F. (1992). Rent generation during the transition to a managed fishery: the case of the New Zealand ITQ System. *Marine Resource Economics, 7*(4), 229–248.

McKelvey, R. (1985). Decentralized regulation of a common property renewable resource industry with irreversible investment. *Journal of Environmental Economics and Management, 12*, 287–307.

Pope, J. G. (1973). An investigation into the effects of variable rates of the exploitation of fishery resources. In M. S. Bartlett & R. W. Hiorns (Eds.) *The mathematical theory of the dynamics of biological populations*. New York: Academic.

Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Research Board of Canada, 11*, 559–623.

Sanchiro, J. N., & Wilen, J. E. (2007). Global marine fisheries resources: status and prospects. *International Journal of Global Environmental Issues, 7*(2/3), 106–118.

Sandal, L. K., & Steinshamn, S. I. (1997). Optimal steady states and the effects of discounting. *Marine Resource Economics, 12*, 95–105.

Sandal, L. K., & Steinshamn, S. I. (2001). A simplified feedback approach to optimal resource management. *Natural Resource Modeling, 14*, 419–432.

Sandal, L. K., Steinshamn, S. I., & Hoff, A. (2007). Irreversible investments revisited. *Marine Resource Economics, 22*, 255–266.

Schaefer, M. B. (1954). Some aspects of the dynamics of populations important to the management of commercial marine fisheries. *Bulletin of the Inter-American Tropical Tuna Commission, 1*, 25–56.

Schaefer, M. B. (1957). Some considerations of population dynamics and economics in relation to the management of marine fisheries. *Journal of the Fisheries Research Board of Canada, 14*, 669–681.

Scott, A. D. (1955). The fishery: the objectives of sole ownership. *Journal of Political Economy, 63*, 116–124.

Smith, V. (1969). On models of commercial fishing. *Journal of Political Economy, 77*, 181–198.

Warming, J. (1911). Om grundrente af fiskegrunde. *Nationaløkonomisk Tidsskrift, 49*, 499–505 (in Danish).

Absorptive Capacity and Social Capital: Innovation and Environmental Regulation

Arent Greve[1]

Abstract Norwegian paper and pulp mills are subject to strict environmental regulation. The mills conduct research and development for reducing pollution. Absorptive capacity indicates their competence. Firms are part of a social network of connections with external actors that include other paper and pulp mills, suppliers, customers, research institutes, and universities that help them in developing technologies. These relations represent their social capital. Some firms have access to more and better resources than other firms. Measuring firms' absorptive capacity and access to social capital, we analyze their success in reducing pollution levels. There is a strong interaction effect between absorptive capacity and social capital. The effectiveness of social capital depends on absorptive capacity, and absorptive capacity is not effective without social capital.

1 Introduction

The paper and pulp industry was infamous for its pollution of the environment. If nothing is done about effluents, pulping processes can kill all life in the rivers, lakes, and fjords into which their waste flows. One of our respondents had by 1975 killed a whole fjord, but by 1998, 99% of life had returned to the fjord. The investments were huge, but very profitable, so that in 2–3 years they had been paid back. The Norwegian government started regulating pollution in the 1970s. During the first two decades, neither regulators nor firms knew how to apply or develop technology to reduce pollution levels. The firms realized that they had to make

[1] In collaboration with Rolf Golombek, The Frisch center, The University of Oslo, and Ken Harris, Melior.

A. Greve
Department of Strategy and Management, Norwegian School of Economics and Business Administration (NHH), Breiviksveien 40, 5045 Bergen, Norway
and
Fakultet for økonomi og samfunnskunnskap, Universitetet i Agder, Kristiansand, Norway
e-mail: arent.greve@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_21,
© Springer-Verlag Berlin Heidelberg 2010

fundamental changes. Most of the firms in this study had to radically change their core technology. Most efforts before 1990 had marginal results, but during the 1990s the firms made progress. By 1998, all of them had installed new process technology or cleaning plants. This study shows how ten firms innovated and adopted pollution reduction technologies that not only reduced pollution to low levels, but also became profitable themselves through better resource utilization. We analyze how the use of internal and external knowledge helped in developing and using pollution reducing technologies. However, success was not evenly distributed; some firms are extremely successful, while others have failed miserably.

Most theories look at firms' knowledge base as their prime resource for innovations (Wernerfeldt, 1984, 1995; Conner and Prahalad 1996). Cohen and Levinthal (1990) define absorptive capacity as a firm's ability to innovate. Innovation requires use of internal as well as external resources; knowledge is one of the most crucial resources to succeed (Fabrizio 2009). In this paper, we look at how firms use their internal knowledge to access knowledge from external contacts to help them develop and run technologies that reduce pollution. To reduce pollution, firms use complex technologies, which are defined as technologies that consist of components requiring interdisciplinary knowledge at a high level.

We first review literature on innovation studies and how absorptive capacity, social networks, and alliances of firms are instrumental in creating innovations. We combine these fields of research and propose that the firms' success in developing technology depends on the combination of absorptive capacity and their access to social capital.

1.1 Previous Research on Absorptive Capacity and Social Networks

Cohen and Levinthal (1990) measure absorptive capacity as research and development (R&D) expenses, which reflect the extent to which firms develop internal knowledge to take advantage of external knowledge. Research on effects of absorptive capacity has been done at several levels. A few studies have been carried out on transfer of technology between organizations (Von Hippel 1988; Levinson and Asahi 1995). These studies demonstrate that during technology transfer, organizations may learn and develop new products or processes, but their ability to successfully innovate depends on the companies' absorptive capacity (Powell 1998).

Studies of technology development have also applied a social network approach. Technologically advanced innovations require efforts from several persons from several firms. Saxenian (1994) compares Route 128 with Silicon Valley and finds that access to networks is crucial for successful innovations. Studies of the biotechnology industry show that few firms can mobilize sufficient internal resources to accomplish R&D, patenting, clinical trials, governmental approval, production, and commercialization on their own. They enter strategic alliances that help them grow and prosper (Powell et al. 1999).

Many studies take the internal capacity for doing R&D as granted and concentrate on effects of external relations. However, Bougrain and Haudeville (2002) show that external collaborations do not increase the chances of success unless the firms are prepared; they need to develop and organize their own capacity for doing R&D, see also Niosi (2003). Successfully innovating firms link their capabilities to complementary resources of other firms (Teece 1987). These collaborations can benefit both parties in an exchange of knowledge (Von Hippel 1988). R&D has become more interorganizational during the last few decades; most projects, particularly those dealing with complex technologies, involve several firms and research institutions (Powell et al. 1996, 1999; Powell 1998; Greve and Salaff 2001). Most studies seem to have studied either firms' absorptive capacity, or their formal external relations, which in most cases are defined as formal alliances. Few studies combine these factors and include informal relations; notable exceptions are Bouty (2000), Kreiner and Schultz (1993), Soekijad and Andriessen (2003), and Tsai (2001).

This study looks at firms with a large range in resources. Some firms have built R&D competence over years of developing production technologies; others have done less or almost nothing in this area. All of the firms use external relations in their work involving reduction of pollution. This study analyzes effects of using both internal and external resources to solve a shared problem within one industry. This makes comparisons across firms more valid by eliminating heterogeneity in the sample. Compared to other network studies, we use direct measures based on interviews, instead of indirect measures based on public data about contracts and collaborations.

1.2 Theory: Absorptive Capacity and Social Capital

1.2.1 Absorptive Capacity

Absorptive capacity refers to how a firm's knowledge influences its ability to develop products or production processes. Knowledge is usually embedded in the professions that develop or manage the firm's technology (Scott 2001, 2008). The original operationalization of absorptive capacity uses a firm's R&D spending as an indicator of its accumulation of knowledge that can be applied to change. Later studies do not question this operationalization; however, there are a few exceptions. Lane and Lubatkin (1998) introduce the term relative absorptive capacity defined as a dyadic relation between firms, the most important being familiarity with problems and a match of the specific knowledge of the partners (Kim and Kogut 1996). To learn, a firm has to build new knowledge on existing and related knowledge involving its activities (Mowery et al. 1996). Lane and Lubatkin (1998) stress the value of interactive learning, as this is the only way firms can learn and develop knowledge to solve problems. Ahuja and Katila (2001) studying knowledge transfer in acquisitions, report similar findings, and they stress the importance of absolute size of the knowledge base that enters a new relation.

Levinson and Asahi (1995) illustrate through case studies that absorptive capacity enhances interfirm technology transfer. These studies point in the direction of connecting a firm's ability to learn from other firms to relations with specific other firms that to a large extent match their prior knowledge (Mowery et al. 1998). This anchors absorptive capacity to prior investments in technology and to the specific relations that firms use to innovate. Thus, the development of absorptive capacity not only depends on a firm's knowledge base, but also on its ability to interact with other knowledgeable firms and institutions that match its own knowledge base (Fabrizio 2009). Absorptive capacity is seen as a required component for successful innovations. We expect that firms with high absorptive capacity would be more successful in their R&D and innovation to reduce pollution.

Therefore, we may not expect absorptive capacity to work unless the firms also have useful external relations that help them innovate (Cockburn and Henderson 1998). Most prior studies do not control for external relations, or they focus on either absorptive capacity or on the effects of external relations. Many findings may have been confounded by unmeasured variables. Another problem is related to the difficulty of the problems an organization faces. Most organizations solve problems without consulting external experts; however, some problems may be too complex to solve on their own, and they may need external help (Soekijad and Andriessen 2003). Therefore, we suggest that firms with absorptive capacity but without external support may not be able to reduce pollution when complex problems are present.

1.2.2 Social Capital

Firms and their employees are embedded in networks of social relations (Granovetter 1985; Gabbay and Leenders 1999). The instrumental aspects of social capital are clear and Lin (2001, p. 19) defines social capital as: "investment in social relations with expected returns in the marketplace." By investing in relations, firms can accumulate social capital. This definition distinguishes social relations that people use for achieving goals, and this is in accord with most writings on social capital (Bourdieu 1983, 1986; Coleman 1988; Burt 1992; Portes 1998). Social capital is part of a larger social structure. It is a common good that those with the same relations share. However, even shared resources have a different value depending on the receiver's need. Social structures contain economic, professional, and personal relations. These relations extend across organizations, such as professional networks and colleagues from earlier jobs (Greve and Salaff 2001).

Social capital is useful where common understanding already exists, facilitating sharing of knowledge and coordinating problem solving (Lane and Lubatkin 1998; Ahuja and Katila 2001). Drawing on external contacts enables engineers to mobilize complementary resources to supplement their firm's assets and solve problems they would not be able to solve on their own. Von Hippel (1988) shows how small steel mills helped one another overcoming production problems by letting key employees visit other plants and discuss these problems. Although competitors shared solutions

to production problems and thereby shortened lead times of first movers, such exchanges strengthened the community of firms, and guaranteed access to other firms whenever a problem occurred within any firm. We observe the same open access to knowledge in the paper and pulp industry. Pollution control in paper and pulp processes requires knowledge in organic chemistry in addition to paper and pulping processes. To solve complex problems the mills need to coordinate a set of different areas of knowledge. Part of it is known within the firm, and part of it is knowledge that exists within the industry or at the research frontier in universities and research institutes. Some of this knowledge is tested and verified, whereas other theories have not yet been verified. Engineers have to combine knowledge with different degrees of certainty; they need to know about and search for evidence, and conduct experiments in areas where the firms and the community lack verified knowledge.

However, without extensive internal knowledge it is hard to see how external help can solve complex technical problems of firms. We may expect that access to large social capital may not have an impact unless there is also vast internal knowledge (Ahuja 2000; Niosi 2003). Several studies on networks and alliances did not control for the firms' level of internal knowledge, or absorptive capacity. Therefore, some of the results may have been confounded by unmeasured variables. We would not expect firms without absorptive capacity but with large social capital to be able to solve complex problems and reduce pollution.

1.3 Combining Absorptive Capacity and Social Capital

If possession of high absorptive capacity attracts other firms as discussion partners or collaborators, such reciprocity should benefit both partners by augmenting their knowledge through cooperation (Rogers and Larsen 1984; Von Hippel 1988; Saxenian 1994). Cockburn and Henderson (1998) show that in the pharmaceutical industry, building absorptive capacity by only doing in-house R&D, does not contribute to performance measured as patents. However, joint research with universities contributed to productivity. Fabrizio (2009) found that joint research with universities and doing basic research in-house enabled biotechnology firms to innovate faster with higher quality. Thus, social capital brings mutual development of knowledge. For any single firm, this kind of reciprocity also has system wide effects. Network clusters whose participants have high absorptive capacity, have high quality. We do not expect these processes to occur where absorptive capacity is not balanced by social capital. Tsai (2001) reports that in a divisionalized firm, innovation and division performance depend on an interaction effect between interdivision networks and absorptive capacity.

Hypothesis 1: An interaction effect between absorptive capacity and social capital reduces pollution levels. We expect firms with high absorptive capacity and large social capital to reduce pollution more efficiently than firms with less of these resources (negative effect on pollution levels).

2 Methods

2.1 Data and Respondents

2.1.1 Research Design

The regulatory agency (SFT) has continuously measured pollution during the last three decades. The real progress in dealing with pollution started around 1990. We aimed at a causal design measuring absorptive capacity and social capital during 5–8 years preceding 1995 (for year of the data source, see below). We mapped social networks in January/February 1998, with a mean duration of ties at 8.5 years. Our measurement of pollution levels dates from 1998, giving time for investments to take effect.

2.1.2 The Study Setting

Pollution levels are regulated individually for each firm in Norway. SFT issues permissions for pollution levels to each firm, then reviews the emissions over a 5 year period, and imposes stricter regulations later. This individual treatment sets firms on their own trajectories with independent rates of change, depending on how good they are to solve problems. This makes an excellent setting to do research on technology development.

The research design allows for looking at effects of the independent variables over time, as they are recorded for a time period of 5–8 years before the measures on the dependent variable. The data collection started in 1995 and ended early 1998. The dependent variable was obtained in 1999, and reflects the pollution levels during 1998.

2.1.3 Data Sources and Measurements

Our methods combine single firm case studies and data from five sources. (1) Measures of the dependent variable come from records from SFT of emissions from the mills. We have data on two components of water pollution. Most water pollution in the pulping process is caused by COD (dissolved organic substances) and suspended solids (SS). These two pollutants consume oxygen from the water during decomposition. In severe cases whole lakes or fjords can become oxygen depleted, and all life forms perish. We calculate our two measures of water pollution, COD% and SS%, as the pollution percentage of total production measured in tons paper or pulp. In the 1970s, mills wasted up to 55% of their input. Today, the best firms aim for zero emissions. The combined SS and COD outputs in our sample vary between 0.82% and 20%.

The measures of the independent variables also come from (2) three decades of archives of correspondence between the mills and SFT; and (3) extensive interviews

at all the mills and some of their most important network contacts. These sources (2 & 3) gave insights into who participated, and what they did. (4) The Association of Pulp and Paper Engineers provided data on employees, their jobs, and education levels. To measure absorptive capacity in each firm, we count the persons within the professions of the industry, who were holding higher degrees, employed during the previous 5 years and working on production technology and environmental issues. Among the 4,430 employees in the ten firms, there are ten PhDs in five firms besides 141 MSc and MPhil degrees. Absorptive capacity is a firm level property, but individuals have the underlying knowledge to process nonroutine information. Measuring R&D costs is a problematic measure for several reasons. Our study firms do not have comparable ways to allocate costs for R&D and other activities. Investments in R&D may also have variations in the outcome relative to spending. Counting people involved in reducing pollution has the advantage of comparing similar types of competencies across firms (Conner and Prahalad 1996), thus giving a microlevel measurement that underlies the macro level concept of absorptive capacity (Foss et al. 2010).

(5) Interview data on the social networks of the firms to assess their social capital. To collect the network data, we interviewed key actors in all the firms, one major technology supplier, and two research organizations. For each firm in the network we asked (a) when they established the connection; (b) if the connection was still used; and (c) how frequently they were in contact. Most relations are long-term; mean duration is 8.5 years. We collected the network data during February and March 1998. We use information centrality measures in sociocentric networks to measure access to social capital, because respondents' evaluation of their contacts creates problematic comparisons across respondents (Snijders 1999). What one respondent experiences as a valuable contribution to pollution reduction may not be useful to another firm. The total number of firms or institutions in the network is 126 from eight countries. From the interviews, we constructed edge lists to prepare sociocentered network matrices (Borgatti et al. 2002). Information centrality takes network effects into account. It is a measure derived from information theory and flow betweenness centrality (Stephenson and Zelen 1989). It indicates the amount of information at each node in the network, calculated from the flow of communications through all paths of a network from all other nodes. Information is reduced by a loss function as it passes through nodes, so that a long path has more loss than a short path. Information centrality assesses the knowledge that can accumulate to any actor due to their position in the network, calculating the effects of both direct and indirect links. Information centrality provides a full picture of the distribution of social capital. To be able to infer causality, both independent variables, absorptive capacity and social capital, are recorded with a time lag to the dependent variable.

2.1.4 Respondents

Our sample represents the majority of the big producers, and hence the historically big polluters. The ten firms selected for this study represent the full range of technologies in the Norwegian pulp and paper industry. There are eight integrated

Table 1 Means, standard deviations, and correlations

Variable	Mean	SD	Correlations Pollution	Absorptive capacity	Social capital
Pollution	6.01	7.19	1.00		
Absorptive capacity	17.10	10.30	0.002	1.00	
Social capital	1.67	0.11	−0.122	−0.040	1.00

producers of both pulp and paper. One firm solely processed paper and one only pulp. These ten firms represent all available pulping processes. Statistics Norway provided us with detailed data and key statistics about the industry; these support the representativeness of the production and main pollution of the industry. Apart from our mills, 22 small paper mills without pollution problems are not part of our study. These small mills run simple and relatively clean production processes. Our sample does not include 13 small mills doing only simple mechanical pulping. We included one single paper mill in our sample; it is typical of its kind in Norway and gave us a chance to study how they were different from the integrated pulp and paper producers. Our sample has more than 80% of the pulp and paper production in Norway. Table 1 displays the distributions and correlations of the variables.

2.2 Regression Analysis

The hypothesis is stated as a relationship with an interaction term, testing for a negative value of b_3:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2 + e,$$
Y = pollution levels relative to production,
X_1 = absorptive capacity,
X_2 = social capital.

In models with direct effects, interpretation of the regression coefficients is straight forward; b_i show the relative increase (decrease) in Y given a change of one unit of each independent variable. Introducing an interaction term changes the interpretation of the direct effects into conditional relationships for the value of X_1 or X_2 when the other regressor is set at 0. The coefficients are no longer tests of main effects (Jaccard and Turrisi 2003). We use mean centering for both variables, which is a recoding of the variables by subtracting their mean. This does not affect statistical tests of significance nor correlations; however, the coefficients b_i $(i = 1, 2)$ are affected but not b_3. The coefficients B_1 and B_2 reflect simple effects, a change in Y when the other variable is set at its mean (Jaccard and Turrisi 2003).

The problems of statistics with small samples are related to external validity and getting statistically significant estimates. In addition to the t-tests for significance of the regression parameters, we also present confidence intervals and power analyses. The latter is a function of a noncentrality estimate that has been adjusted to remove positive bias in the original estimate. It is the probability of achieving a significance level of alpha ≤ 0.05 in a similar sample (Wright and O'Brien 1988).

We apply Ordinary Least Squares estimates using effect screening procedures. This is an appropriate method if we have a large number of effects compared to observations. This method corrects for scaling and correlations among the estimates. The method calculates correlations among estimates and applies a recoding to make them uncorrelated with equal variances. Orthogonally coded estimates avoid regression errors due to correlations among predictors and parameter estimates. We use these estimates to identify variables with sizable effects (Lenth 1989; Berry 1993; Sall et al. 2007). The column labeled "Orthogonally Coded" shows the new estimates of each parameter in the regression with t-test statistics. The parameters are scale invariant and can be compared directly to each other. They are also true estimates as all variables are transformed to be independent.

To resolve which effects are important, we compute posterior probabilities using a Bayesian approach (Box and Meyer 1986). The Bayes estimate is done with respect to normalized estimates (the Orthogonal t-test). The Bayesian approach avoids the double negative of classical statistics which assumes that the complement (the 0-hypothesis) is true, and then tests if the data are inconsistent with this assumption. The classical type of inference depends on sampling and representativeness. The test informs us of nothing but the probability of finding the present distributions in a random data set, which means we do not really know if the alternative hypotheses are true. Bayesian statistics build on prior knowledge and the hypotheses of a stated relationship between variables. It takes the value of collected data and computes the probability of having true estimates regardless of why the data set is limited to the one at hand or a theoretically different data set (Iversen 1984). Thus, the Bayesian approach produces a more robust test of hypotheses.

The issue of external validity depends on representativeness of respondents. How representative is the population of Norwegian paper and pulp mills to the rest of the world, or to other types of innovators struggling with complex innovations? In this case, our sample is representative of the Norwegian paper and pulp production. The remaining question is whether the results can be generalized outside of this setting. This is a question that requires empirical research in settings with complex technologies.

3 Results

3.1 The respondents' History of Pollution Reduction

Because of the small sample size, we discuss the mills and the results of the qualitative part of the study first, and then turn to the tests of the hypothesis. Our interviews and the correspondence between SFT and the firms helped us understand how these firms negotiated pollution levels and adapted to regulations. Some of the people we talked to had been working in this area all their careers, the oldest starting from the mid 1960s. They gave us detailed accounts of their firms. Over the 25–30 year period of our study, the firms have achieved huge reductions of pollution. This was

not a smooth voyage. Between 1970 and 1990, the early period of regulation, few knew how to reduce emissions. Since there was little common knowledge of technologies to reduce pollution in Norway, SFT looked at Swedish companies to learn from their practice. SFT quite often had specific suggestions to what solutions the companies should use. Later, as the firms gained more knowledge, which to a large extent was process specific, they were more able than SFT to design solutions. Most of the negotiations were on how much they were allowed to pollute and scheduling of reductions. Their efforts and abilities developed in various directions and with different results. In the late 1980s, pressures from their customers induced more changes in production. Examples are reduction and elimination of chlorine based bleaching, recycling paper in the production process, and outcries against use of lumber from rain forests. The pressure from their customers coincided with their increased ability to deal with pollution reduction.

Most firms, with a few exceptions, resisted regulation until 1990; almost all the firms have during the 1990s changed their strategies to a proactive behavior. The mills systematically work on pollution reduction independently of the regulations, and some of them reduce their pollution at a higher rate than SFT demands. They have markedly built up their ability to deal with pollution problems during the 1990s, improving their pollution levels continuously. There are quite substantial differences in the development of pollution levels. During the regulation period most of the firms changed process technology or invested in end-of-pipe cleaning plants often combined with recycling or energy production.

3.2 Test of the Hypothesis: Absorptive Capacity and Social Capital

Table 2 displays the results of the regression equations. B is the nonstandardized regression coefficient, significance tests are t-tests, with the exception of the whole model test, which is an F test.

The regressions show that neither absorptive capacity nor social capital has independent, direct effects on pollution levels, see Model 1 in Table 2. The model supports the hypothesis with a significant interaction effect between absorptive capacity and social capital, $R^2 = 0.82$ (adjusted $R^2 = 0.74$), $B_3 = -4.68$, and power test $= 0.99$ (adjusted power $= 0.93$). Social capital is only useful when absorptive capacity is at a high level. Consequently, absorptive capacity is efficient only when combined with social capital. The column "Orthogonally Coded" shows the final estimates after effect screening. Finally, we produce a Bayes estimate of conditional probabilities. Table 3 shows the results. This confirms that the interaction effect, B_3, is the only effect in the model, with a posterior probability $= 0.97$.

In the paper and pulp industry, there are open discussions of production and pollution related problems among engineers in different firms, and information is available to anyone. However, to take advantage of this openness requires absorptive capacity. In dealing with complex pollution issues high absorptive capacity is not of great value unless social capital is present. Firms without high absorptive capacity cannot take advantage of social capital, nor is absorptive capacity by itself enough to

Table 2 Regressions with pollution as the dependent variable

Variable	Model 1 B Estimate (SD)	Model 2 B Estimate (SD)	Confidence intervals		B Orthogonally coded estimate
Intercept	19.45	37.93 (19.38)	−9.50	85.35	6.01**
Absorptive capacity	−0.00 (0.26)	−0.20 (0.13)	−0.51	0.11	0.01
Social capital	−8.05 (24.71)	−17.24 (11.44)	−45.22 −6.86	10.74 −2.50	−0.83 −6.14**
Social capital × absorptive capacity (interaction)		−4.68** (0.89)			
R^2	0.01	0.82	Adjusted R^2		0.74
F (model)	0.05	9.33*			
n	10	10			

Residual (e) = 0.00 SD = 3.02, normal distribution, Shapiro-Wilk W, $p = 0.95$, $p < W = 0.68$, two-tailed *t*-tests

* ≤ 0.05

** ≤ 0.01

Table 3 Bayes test of the regression model for pollution

Variable	Estimate, orthog. *t*-test	Prior probabilities	Posterior probabilities
Absorptive capacity	0.01	0.20	0.02
Social capital	0.71	0.20	0.03
Absorptive capacity × social capital (interaction)	5.24	0.20	0.97

Posterior probability that sample is random (uncontaminated): 0.026

solve pollution problems. Firms need both. Knowledge to reduce pollution levels is partly localized within firms in the form of knowledge of their production systems, and partly located in the environment as specialized knowledge about technologies for reducing pollution. These regressions clearly reveal the interaction between absorptive capacity and social capital.

4 Discussion and Conclusion

To handle the complex problem solving entailed in pollution reduction, pulp and paper mills share specialized knowledge. To design technologies and optimize processes, they need R&D, which includes coordinating specialists, conducting experiments, and organizing projects. Success depends on their absorptive capacity and access to social capital.

Firms build absorptive capacity by investing in technology and people. Because of the complexity of reducing pollution, few possess all the relevant knowledge

and they need to mobilize complementary knowledge. The firms' knowledge is embedded in their production equipment, which may last up to half a century. The investments in technology create path dependence in knowledge requirements and specialization of competence. Doing R&D and solving problems to reduce pollution help the firms build absorptive capacity.

Employees use their social networks to find efficient complementary knowledge, the choice of which is not always obvious. The social network analysis of the ten firms in advice networks on environmental problems demonstrates how knowledge is distributed among a large number of actors; the total network has 126 organizations.

It takes a long time to develop absorptive capacity to revamp production processes, install cleaning plants, and run them. This learning process took place over two to three decades, aided by the long careers of engineers in these firms and in the industry. At the same time that these firms built absorptive capacity, they also built their social networks that are crucial in helping with innovations and problem solving.

As firms discover that they cannot handle their problems, they turn to outside experts, and if they are not able to solve their problems, they continue searching, drawing on more and more external relations. This was the case for the one firm with the highest number of direct relations in our study, but low in absorptive capacity. Their extensive use of social capital was like a cry for help. In the absence of absorptive capacity their numerous relations did not help them. Alas, they turned to a set of firms mainly located among chemical suppliers and consultants that might not have been a good choice. Another firm, highest in absorptive capacity, used external networks to a lesser extent than other firms. They seemed less focused on pollution reduction than other firms, and despite pioneering a few pollution reduction processes they were content with their pollution levels and did not put much effort into improvements.

The interaction between absorptive capacity and social capital reveals that high levels of absorptive capacity together with good access to social capital have the largest effect on pollution levels. Pollution control is complex, it requires multidisciplinary expertise; not even the largest firms know how to deal with the problems on their own. Firms with low absorptive capacity seem not to be able to take advantage of their social capital.

Several studies have been done on absorptive capacity and on social capital, however, few if any has tried to separate the effects of each or to compare the combined effects. Tsai (2001) is a notable exception, testing for internal networks in a divisionalized firm. He finds both direct and interaction effects of absorptive capacity and social capital. In contrast, this study shows that absorptive capacity has primarily effects in combination with social capital, which is contingent on the complexity of the problems and technologies. By demonstrating that absorptive capacity and social capital work in interaction, we open the way for further analysis of complex innovations. This augments Cohen and Levinthal's (1990) original definition of absorptive capacity, which assumes that absorptive capacity is used to take advantage of external information, which remained undefined as if everybody had the same

access to external knowledge. This study shows that access to external contacts is not only unevenly distributed, but also that the ability of the firms to take advantage of what information they get is unevenly distributed. Few firms can accomplish successful innovations to reduce pollution by working alone; being able to draw on their social capital, firms can take advantage of environmental changes that require a sophisticated combination of specialized knowledge.

The main limitation of this study is the sample size. However, it includes about 80% of the volume of the Norwegian paper and pulp industry, and as such can be seen as representative of this industry. One remaining question is if the findings in this study can be applied to similar firms in other countries, and across different industries. This is an empirical question that requires more studies. To achieve a comparable base, it is important to concentrate similar studies on complex technologies. An important assumption in the theory behind this study is that problem solving involves more than one discipline. We would expect that firms with high absorptive capacity can more easily solve problems that arise within the domain of a single discipline. As complexity and need for multidisciplinary knowledge arise, the social capital component becomes more important. The problems that the firms in this study worked on required not only developing new technologies, but also developing or absorbing new fundamental knowledge within disciplines that previously had not been deployed within this industry. Thus, we see the role of social capital as increasingly important with increasing complexity in problem solving. Studies that can achieve variation in complexity would be able to test this assumption. Only one firm had a less complex technology than the others, and it has low absorptive capacity as well as low social capital scores. It scores low on total pollution, ranked as number four because of its low COD levels; however, the SS level is fairly high, and for that type of pollution it ranks as number eight (high rank is negative). This indicates that level of complexity is an important variable that should be tested in future studies to moderate the effects of absorptive capacity and social capital. The source of social capital matters. The most successful of our respondents had contacts to the research frontier in universities. Other, less successful firms consumed second-hand knowledge from consultants, which apparently was not efficient. This is another finding that should be followed up.

Acknowledgments I am grateful to Woody Powell, Simon Rodan, and other participants at the SCANCOR seminar for helpful comments, and also Joel Baum, Brian Silverman, Kristina Dahlin, and other participants at the research seminar at Rotman School of Business, Toronto. Howard Aldrich, Henrich Greve, Arne Kalleberg, and two reviewers for this book have given valuable comments.

References

Ahuja, G. (2000). The duality of collaboration: Inducements and opportunities in the formation of interfirm linkages. *Strategic Management Journal, 21*, 317–343

Ahuja, G., & Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study. *Strategic Management Journal, 22*(3), 197–220

Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage Publications

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *UCINET 6.79*. Natick: Analytic Technologies

Bougrain, F., & Haudeville, B. (2002). Innovation, collaboration and SMEs internal research capacities. *Research Policy, 31*, 735–747

Bourdieu, P. (1983). Economic capital, cultural capital, social capital. (Ökonomisches Kapital, kulturelles Kapital, soziales Kapital). *Soziale Welt, Sonderheft, 2*, 183–198

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). Westport, CT: Greenwood Press

Bouty, I. (2000). Interpersonal and interaction influences on informal resource exchanges between R&D researchers across organizational boundaries. *Academy of Management Journal, 43*(1), 50–65

Box, G. E. P., & Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics, 28*(1), 11–18

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press

Cockburn, I. M., & Henderson, R. M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *The Journal of Industrial Economics, XLVI*, 157–182

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 15*, 128–152

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94 Supplement*, S95–S120

Conner, K. R., & Prahalad, C. K. (1996). A resource-based theory of the firm: Knowledge versus opportunism. *Organization Science, 7*(5), 477–501

Fabrizio, K. R. (2009). Absorptive capacity and the search for innovation. *Research Policy, 38*(2), 255–267

Foss, N. J. Husted, K., & Michailova, S. (2010). Governing knowledge sharing in organisations: Levels of analysis, mechanisms, and research directions. *Journal of Management Studies*, *47*(3), 455–482

Gabbay, S. M., & Leenders, R. T. A. J. (1999). CSC: The structure of advantage and disadvantage. In R. T. A. J. Leenders & S. M. Gabbay (Eds.), *Corporate social capital and liability* (pp. 1–14). Boston: Kluwer Academic Press

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology, 91*(3), 481–510

Greve, A., & Salaff, J. W. (2001). The development of corporate social capital in complex innovation processes. In S. M. Gabbay & R. T. A. J. Leenders (Eds.), *Research in the sociology of organizations (Vol. 18): Social capital of organizations* (pp. 107–134). Amsterdam: JAI Press

Iversen, G. R. (1984). *Bayesian statistical inference*. Newbury Park, CA: Sage Publications

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression*. Thousand Oaks, CA: Sage Publications

Kim, D.-J., & Kogut, B. (1996). Technological platforms and diversifications. *Organization Science, 17*, 283–301

Kreiner, K., & Schultz, M. (1993). Informal collaboration in R&D: The formation of networks across organizations. *Organization Studies, 14*(2), 189–209

Lane, P. J., & Lubatkin, M. (1998). Relative absorptive capacity and interorganizational learning. *Strategic Management Journal, 19*, 461–477

Lenth, R. V. (1989). Quick and easy analysis of unreplicated fractional factorials. *Technometrics, 31*(4), 469–473

Levinson, N. S., & Asahi, M. (1995). Cross-national alliances and interorganizational learning. Organizational Dynamics, 24, 50–63

Lin, N. (2001). *Social capital: A theory of social structure and action*. Cambridge, UK: Cambridge University Press

Mowery, D. C., Oxley, J. E., & Silverman, B. R. (1998). Technological overlap and interfirm cooperation: Implications for the resource-based view of the firm. *Research Policy, 27*, 507–523

Mowery, D. C., Oxley, J. E., & Silverman, B. S. (1996). Strategic alliance and interfirm knowledge transfer. *Strategic Management Journal*, 17(Winter Special Issue), 77–91

Niosi, J. (2003). Alliances are not enough explaining rapid growth in biotechnology firms. *Research Policy, 32*, 737–750

Portes, A. (1998). Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology, 22*, 1–24

Powell, W. W. (1998). Learning from collaboration: Knowledge and networks in the biotechnology and pharmaceutical industries. *California Management Review, 40*, 228–240

Powell, W. W., Koput, K. W., & Smith-Doer, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly, 41*(1), 116

Powell, W. W., Koput, K. W., Smith-Doerr, L., & Owen-Smith, J. (1999). Network position and firm performance: Organizational returns to collaboration. In S. Andrews & D. Knoke (Eds.), *Research in the sociology of organizations* (Vol. 16, pp. 129–159). Greenwich, CT: JAI Press

Rogers, E. M., & Larsen, J. K. (1984). *Silicon Valley fever: Growth of high-technology culture*. New York: Basic Books

Sall, J., Creihton, L., Lehman, A., et al. (2007). *JMP start statistics: A guide to statistics and data analysis using JMP and JMP-IN software*. Cary, NC: SAS Institute

Saxenian, A. (1994). *Regional advantage: Culture and competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press

Scott, W. R. (2001). *Institutions and organizations*, 2nd edn. Thousand Oaks, CA: Sage Publications

Scott, W. R. (2008). Lords of the dance: Professionals as institutional agents. *Organization Studies, 29*(2), 219–238

Snijders, T. A. B. (1999). *Prologue to the measurement of social capital. The Tocqueville Review, XX*(1), 27–44

Soekijad, M., & Andriessen, E. (2003). Conditions for knowledge sharing in competitive alliances. *European Management Journal, 21*, 578–587

Stephenson, K., & Zelen, M. (1989). Rethinking centrality: Methods and applications. *Social Networks, 11*(1), 1–37

Teece, D. J. (1987). Profiting from technological innovation: Implications for integration, collaboration, licensing, and public policy. In D. J. Teece (Ed.), *The competitive challenge: Strategies for industrial innovation and renewal*. Cambridge, MA: Ballinger Publishing Company

Tsai, W. (2001). Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance. *Academy of Management Journal, 44*(5), 996–1004

Von Hippel, E. (1988). *The sources of innovation*. Oxford: New York University Press

Wernerfeldt, B. (1984). A resource-based view of the firm. *Strategic Management Journal, 14*, 4–12

Wernerfeldt, B. (1995). The resource-based view of the firm: Ten years after. *Strategic Management Journal, 16*, 171–174

Wright, S. P., & O'Brien, R. G. (1988). *Power analysis in an enhanced GLM procedure: What it might look like*. Paper presented at the Proceedings of the SAS Users Group International Conference, Cary, NC

Issues in Collaborative Logistics

Sophie D'Amours and Mikael Rönnqvist

Abstract Collaborative logistics is becoming more important in today's industry. This is driven by improved economic and environmental efficiency through collaborative planning supporting resources sharing and new business models implementation. This paper presents a survey of contributions to the field of collaborative logistics. It first describes current opportunities in collaborative planning. It then discusses important issues related to building the coalition, sharing resources and benefits, as well as related to information and decisions technologies. Business cases are described and used to support the discussion. Finally, questions are raised, opening new paths for researchers in the field.

1 Introduction

In this paper, we will discuss issues related to collaboration between companies when dealing with logistics and transportation. Logistics and transportation are activities that provide many opportunities for collaboration between companies. This collaboration, either through information or resource sharing, aims to reduce the cost of executing the logistics activities, improve service, gain market shares, enhance capacities as well as protect the environment, and mitigate climate change (Simchi-Levi et al. 1999). Collaboration occurs when two or more entities form a coalition and exchange or share resources (including information), with the goal of making decisions or realizing activities that will generate benefits. As illustrated in Fig. 1, collaboration can range from information exchange, joint planning, joint execution, up to strategic alliance (e.g., co-evolution) (D'Amours et al. 2004).

S. D'Amours
FORAC-CIRRELT, Université Laval, QC, Canada
e-mail: sophie.DAmours@forac.ulaval.ca

M. Rönnqvist (✉)
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: mikael.ronnqvist@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_22,
© Springer-Verlag Berlin Heidelberg 2010

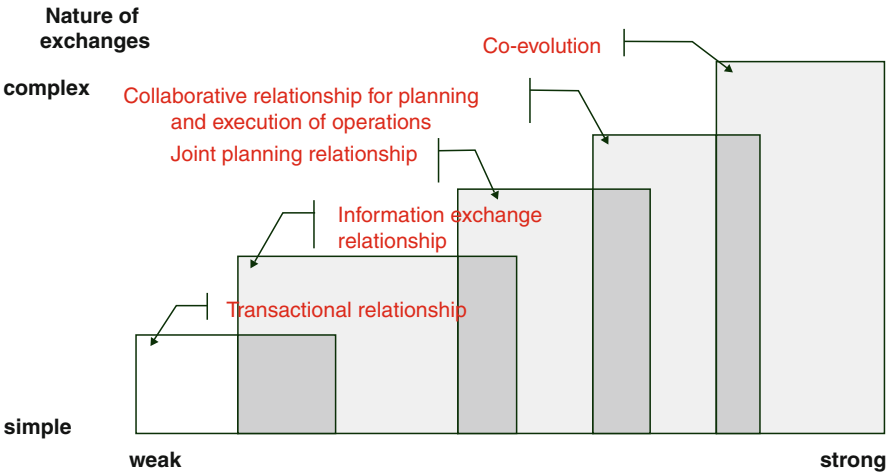


Fig. 1 Different levels of collaboration framework

As collaboration becomes more strategic, typically more resources are involved and sensitive information is shared between the collaborating entities.

In this paper, we discuss important aspects when building a coalition of business entities aiming for collaborative logistics. These are of different natures. After introducing opportunities for collaborative logistics, we will discuss which entities should be part of the coalition and who should decide about this. The second aspect raised is how the coalition will share the benefits gained. The third aspect is the need for specific information and decision technologies. Finally, a discussion will bring up problems related to collaborative logistics.

2 Collaborative Logistics Problems

Logistics deal with moving and storing products as they flow through the supply chain. Efficient logistics planning and execution can provide competitive advantages to the different entities of the supply chain. Figure 2 presents the supply chain as being the network of business entities involved in the production and distribution of products from raw material to delivery. In some cases, the supply chain spans even further, including activities like design and engineering as well as return and recovery of products. Collaboration in logistics based on information exchanged has been identified as one means of reducing the negative impacts of the bullwhip effect, known as the amplification of demand variation going upstream the supply chain (Lee et al. 1997; Moyaux et al. 2004). The bullwhip effect results in higher inventory levels and increased backlogs. Collaboration can considerably reduce these negative effects.

The supply chain entities such as carrier, producer, customer, and third party logistics collaborate in different ways. In terms of transportation, they aim to optimize

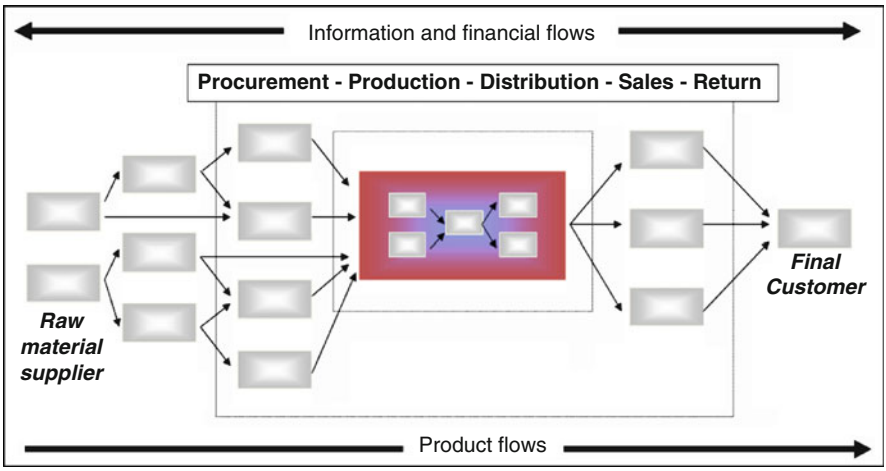


Fig. 2 An illustration of the overall supply chain with its main activities: procurement, production, distribution, sales, and return

the traveling time and load capacity usage. They share information in such a way that the operative pick-up and delivery routing problems capture the benefit of a denser network. They aim to minimize transportation costs, in particular the backhauling costs. Backhauling represents the possibility of combining two transport orders in such a way that the unloaded distance is minimized. The supply chain entities may also collaborate to increase responsiveness and reduce inventory holding costs. In such cases, they share demand and consumption information in a timely manner and use different approaches to synchronize efficiently their activities. The entities may also face large transportation costs and may aim to deploy new infrastructures that will provide them with a competitive advantage over others. Such shared infrastructure could be pipelines (e.g., crude oil and gas), terminals (e.g., forestry), distribution centers or warehouses (e.g., retailing), or transportation modes (integrating, e.g., train, ship, truck in general transportation organizations). Another common practice in collaborative logistics is to share spare parts between different entities.

Collaboration is related to some forms of interdependency. Frayret et al. (2004) have reviewed these forms. They are listed and briefly described in Table 1.

Different cases are reported in the literature. Frisk et al. (2010) discuss collaboration between eight forest companies. These companies wanted to study what the potential savings in tactical monthly planning would be if all companies share all supply and demand. The transportation planning problem can be formulated as a Linear Programming (LP) problem. The potential savings were as much as 14%. About 6% arises from better transportation planning within each company and another 8% comes from the actual cooperation. This is an example of pooled interdependence. Audy et al. (2010) analyses a potential collaboration between four furniture manufacturers. The aim is to optimize collectively the outbound transportation of their products to the US. The case raises interesting issues about benefit

Table 1 Forms of interdependency

Type of relation	Description
Pooled interdependence	Occurs when each part of a system renders a discrete contribution to the whole, while each part is supported by the whole
Producer–consumer relationship or sequential interdependence	Links two manufacturing activities for which the output of one is the input of the other
Reciprocal relationships	Concerns activities whose outputs are the reciprocal inputs of the other activity
Intensive interdependence	Relates to the intrinsic sophistication of activities that are imbedded
Task/sub-task interdependencies	Relates to the decomposition of tasks into sub-tasks
Simultaneity interdependence	Occurs when activities need to be performed, or not, at the same time, such as for meeting scheduling

sharing. This is another example of pooled interdependence. Lehoux et al. (2009) presents a different context which is an example of a producer–consumer relationship. In their paper, a producer and a distributor evaluate the possibility of collaboration. They can follow different logistics approaches (e.g., VMI – Vendor Managed Inventory, CPFR – collaborative planning forecasting and replenishment). In their case, no equilibrium solution can be found; therefore, the partners need to work out some efficient incentives to make it possible.

Collaborative logistics may involve different levels of resource and information sharing between two or many entities as illustrated in Fig. 1. The impact can be strategic. The collaboration can be a voluntary action or imposed by certain policies. Finally, it can bring together business entities which are competitors or suppliers/customers. In such cases, the coalition is said to be, respectively, horizontal and vertical. In some cases, companies acting in complementary sectors of activities may collaborate to develop an extended value proposition. Such networks are defined as diagonal. They are common in the tourist industry, where packages are built through a collection of offers. These perspectives will now be discussed.

2.1 Strategic Collaboration in Logistics

Collaboration can be strategic and therefore imply the sharing of key infrastructures or highly sensitive information. Examples of such collaborations could be the sharing of costly infrastructure. The location and the investment for such infrastructure are considered strategic for the entities involved. Other strategic collaboration relates to defining industry standards; this is the case when entities of a same industry collaborate to define different business standards to improve the interoperability of their systems (e.g., PapiNet – a standard for information exchange in the forest products industry). Strategic collaboration can also imply a long term business contract

and the exchange of demand and capacity information. At the strategic level, it is likely to see entities exchanging a "complete" model of their demand or capacity, permitting the coalition to compute the value of the collaboration and to propose a reasonable sharing strategy (Montreuil et al. 1999; Frisk et al. 2010).

2.2 Operational Collaboration in Logistics

Operational collaboration requires low commitment. An example of such a collaboration could be a web based platform inviting enterprises to share their transportation needs in order to find joint routes that will reduce their transportation costs. Applications can be found in the Swedish and Finnish forest industry. In Eriksson and Rönnqvist (2003), a web-based transportation planning tool is described. Here, a backhauling solution is provided to all transport planners in several regions and companies in order to support manual planning of daily routes.

2.3 Size of the Coalition

Collaboration can bring together two or more entities. In all cases, the need for each entity to improve its logistics is a prerequisite for the collaboration. In a many-to-many context, the design of proper collaboration mechanisms is more difficult, mainly because the exchanges are not bilateral as in a supplier–customer type of collaboration. Externalities play a role in defining when the coalition reaches its optimal size. Some entities may enter with a lot to provide and little to gain, while others will benefit greatly with little to offer.

2.4 Collaboration Driver

Collaboration can emerge from a coalition of voluntary entities that see in the sharing of resource and information ways to improve their logistics (VICS 2004). It can also be imposed by one of the leading entities of the supply chain. For example, WalMart's move to set RFID systems with all major suppliers was done in order to increase the collaboration, but it was imposed on the different entities. Although they are all free to maintain or not their business with WalMart, we define it as an imposed collaboration scheme. Other imposed schemes can be set by public policies. For example, natural resources can be managed by governmental authorities and the allocation rules may impose collaboration between many entities. This is the case in the forestry industry in Canada. In such a situation, the different entities are asked to find harvesting plans which meet the needs of all members of the coalition (Beaudoin et al. 2007).

2.5 Collaboration in Vertical Network

In a context where a supplier and a customer aims for more efficiency in their logistics, they will evaluate the possibility of exchanging more information, plan jointly their activities, and maybe share in the execution of some of the activities. Some responsibilities can also be shifted from one entity to another, so the global efficiency of the coalition is improved as is the case in Vendor Managed Inventory.

We discuss here three types of supplier–customer collaboration models: Vendor Managed Inventory (VMI) and Continuous replenishment and collaborative Planning Forecasting and Replenishment (CPFR).

Under a VMI agreement, the producer is responsible for managing the inventory of its customer. The customer provides the daily consumption to the producer so it can build a production–distribution plan that meets the fixed service level as well as optimizes the usage of its production–distribution resources. The VMI approach showed that a certain amount of collaboration between the supplier and the customer is possible. VMI has contributed positively to enhancing the logistics performance. Danese (2006) reported the benefits gained by the pharmaceutical giant GlaxoSmithKline. Since then, companies have explored new approaches such as Continuous Replenishment models based on carrier capacity or production capacity. The replenishment is structured around a pre-scheduled reservation of capacity. For example, the collaboration may set a one truck per day delivery to the customer. Then, the customer is responsible for setting the mix of products to be on the truck every day. This approach satisfies the needs of the customer over time and reduces the pressure on the producer. The same approach applies with capacity reservation. Finally, the Collaborative Planning, Forecasting and Replenishment (CPFR) business model aims to balance demand and production–distribution capacity up-front in order to define a win-win unique plan for both parties. Cederlund et al. (2007) reported a reduction of 50% of transportation costs and 30% of inventory holding costs at Motorola. Although these collaboration approaches seem to perform well, overcoming many hurdles limits their implementation (Lehoux et al. 2008).

Bilateral collaboration may not reach an equilibrium meaning that no solution is the best for both parties and therefore unless an incentive can be provided, one of the parties will not participate under the defined model. Often, one needs to share the benefits to motivate the others to participate in the collaboration. The design of these incentives is critical for the collaboration (Chen 2003, Cachon 2003).

2.6 Research Questions

Collaborative logistics raises the need for specific methods to support decision-making. The main questions are related to how to build the coalition, sharing policies and information and decision systems.

When dealing with how to build the coalition, the following questions are raised. How will the coalition be built? Which entity or entities will lead the coalition and therefore set its goals? Which entities should be invited to join or leave the coalition?

When dealing with defining the sharing policies, the following questions need to be addressed. How will the value of the collaboration for the whole coalition and for each entity be set? Which sharing method will provide a sustainable solution for the coalition? What should be shared: costs or savings?

Finally, when designing the information and decision technology needed to support the coalition different questions need to be answered. What are the coalition decisions support systems? Which type of electronic platform should be used? Are there any security issues? How to make the systems of the entities work together to transfer information in a timely manner?

The following sections will review current methods and knowledge used to solve these problems.

3 Building Coalitions

We denote by *coalition* a set of stakeholders (e.g., producer, customer, carrier, third party logistics), disposed to share information or resources in order to better realize logistics activities including planning. We denote by *entities* each of these stakeholders and consequently, a coalition must include at least two *entities*. To implement collaborative logistics between these players, we need to build a coalition. From a business point of view, one or a set of the entities will lead in the creation of the coalition. Audy et al. (2007a) have identified six different forms of leadership currently used in coalition building for the specific problem of log supply to mills. We generalize the proposed type of leadership and identify five different models. They are presented in Table 2. In these models, the leader is either one entity which aims to optimize its own objective through some sharing or by a coalition of many that aim to optimize a collective objective.

These models have different objectives and therefore require different methods to support the building of the coalition. The objectives often depend on the leader's attitude. It may behave altruistically, aiming for a fair sharing of the benefits between the coalition members. It can also behave in a more opportunistic way, arguing that

Table 2 Forms of leadership in coalition building for collaborative logistics

Model	Description of the leadership
1	A customer/producer leads the coalition: it aims to minimize its transport costs by finding other customers/producers that can provide a good balance (geographical, volume, and time) between supply and demand
2	A carrier or 3PL leads the coalition: it aims to maximize its profit by a better usage of its carrying capacity
3	A coalition of customers/producers shares the leadership of the coalition: they aim to minimize their transportation costs
4	A coalition of carriers shares the leadership of the coalition: they aim to maximize their profit by a better usage of their joint carrying capacity
5	A coalition of carrier(s) and customer(s)/producers(s) shares the leadership of the coalition: they aim to minimize their transportation costs by using the carrying capacity of the carriers

it is assuming most of the risk and the effort to build the coalition. In such a case, it can propose bilateral offers for each new entity entering the coalition. Depending on the behavior of the leader, the sequence in which the entities are invited into the coalition may have an impact on its benefit. For example, a coalition aims to reduce procurement costs through collaborative routing and supply allocation. The leader is opportunistic and shares costs or savings with one entity at a time, as they enter the coalition. The benefits are shared following a volume ratio. If the costs are shared, then the leader wants to bring in the bigger entities first and then the smaller ones. If the savings are shared, then the inverse logic would be pursued (Audy et al. 2007b).

If we disregard external business considerations, the basic rule of adding an entity p to a coalition c is if the entity p increases the *benefit* of the current coalition c. The *benefit* of a coalition c is defined as the difference between the value of the collaborative plan including all entities in the coalition c, V^c, compared to the sum of the values of the individual plan of each player p in the coalition c, $\sum_{p \in S} V_p$ where S is the set of players in coalition c. In a minimization objective context, the benefit refers generally to the savings whereas in a maximization context they refer to a profit.

A coalition c' will be created if more benefit can be generated by adding entity p' to coalition c. On the other hand, any entity p already in a coalition c who does not contribute to the benefit of this coalition c should be removed.

Although the addition of an entity to a coalition can provide a benefit, it seems that the entities' willingness for the collaboration is tightly linked to the business model of the coalition that is driven by one or several leading players. These leading entities aim at building the coalition in such a way that they will maximize their returns while providing enough incentives to the others to keep them in the coalition.

Evaluating the value of the collaboration for the coalition as well as the entities is fundamental in building the coalition. In logistics, this evaluation is mainly conducted using operational research planning and transportation models, which are sequentially used to solve the entity problem and the different coalitions' problems. Examples can be found in recent papers by Frisk et al. (2010), Audy et al. (2010), and Lehoux et al. (2008).

4 Sharing Principles

4.1 Game Theoretic Background

We will discuss a number of sharing principles once the coalition has been formed and agreed. These have been applied in various industrial settings. Before we describe each of the principles we start by introducing some basic notation used in game theory. We will discuss sharing principles based on cost allocation methods. The motivation is that logistics is more often concerning costs than revenues.

- We have a set of business entities N.
- A *coalition* S is a subset of business entities, i.e., $S \subset N$.

– The *grand coalition* is the set of all entities, i.e., N.
– The cost of a coalition is denoted $c(S)$.

A cost allocation method distributes (or allocates) the total cost of the grand coalition to the entities. This aspect is important as it often is needed to evaluate individual contributions when coalitions are formed, see e.g., Sprumont (1990). Each entity j will be allocated the cost y_j. Since the total cost is to be distributed among the entities, we have

$$\sum_{j \in N} y_j = c(N) \tag{1}$$

A cost allocation which satisfies the above constraint is said to be *efficient*. There are other properties that can be associated with a cost allocation. One property which requires that the entity be not allocated a higher cost than its own cost is called individual rationality. This is simply expressed as

$$y_j \leq c(\{j\}) \tag{2}$$

Another important concept is to ensure that there are no incitements for a coalition to break out and work independently. This implies that the cost allocated to a particular coalition of entities cannot exceed the actual cost of the coalition. There are many potential coalitions and this means that we have one constraint for each possible coalition. This can be expressed as

$$\sum_{j \in S} y_j \leq c(S) \; \forall \; S \subset N \tag{3}$$

Constraint set (1) and (3) define what is called the *core*. Any solution which is feasible with respect to the core is called *stable*. In general, there is no guarantee that there exists a feasible solution to the core. The game is said to be monotone if

$$c(S) \leq c(T), S \subset T \tag{4}$$

This means that if one new entity is included in a coalition, the cost never decreases. The game is said to be *proper* if

$$c(S) + c(T) \geq c(S \cup T), S \cap T = \emptyset \tag{5}$$

This implies that it always profitable (or at least not unprofitable) to form larger coalitions. The properties discussed above are not satisfied for all classes of games. Some may be guaranteed and others not.

For each coalition, S, and a cost allocation, y, we can compute the *excess*

$$e(S, y) = c(S) - \sum_{j \in S} y_j \tag{6}$$

which expresses the difference between the total cost of a coalition and the sum of the costs allocated to its members. For a given cost allocation, the vector of all excesses can be thought of as a measure of how far the cost allocation is from the core. If a cost allocation is not in the core, at least one excess is negative.

4.2 Quantitative Allocation Methods

There exist many allocation rules and we will discuss some that have been used in different applications. We will discuss a limited number of such allocations in this section.

4.2.1 Weighted Costs

A simple and straightforward allocation is to distribute the total cost of the grand coalition among the participants according to a volume or a cost weighted measure. This is expressed by the formula

$$y_j = \frac{c(\{j\})}{\sum_{j \in N} c(\{j\})} c(N) \tag{7}$$

It is intuitive but can often lead to an allocation that does not satisfy, for example, the core conditions.

4.2.2 Separable and Non-Separable Costs

A more advanced method is based on dividing the allocation into two parts. One is associated with a separable cost and the other a non-separable cost. The separable cost or the marginal cost (8) of entity j and the non-separable cost (9) as can be expressed as

$$m_j = c(N) - c(N \backslash \{j\}) \tag{8}$$

$$g_N = c(N) - \sum_{j \in N} m_j \tag{9}$$

Methods based on separable and non-separable costs allocate the costs according to

$$y_j = m_j + \frac{w_j}{\sum_{j \in N} w_j} g_N \tag{10}$$

Depending on which weights are chosen, there are different versions of the method; the two most straightforward methods are the *Equal Charge Method* that distributes

the non-separable cost equally and the *Alternative Cost Avoided Method* that uses the weights $w_j = c(\{j\}) - m_j$, expressing savings that are made for each participant by joining the grand coalition instead of operating alone. These allocations satisfy the efficiency and symmetry properties. However, they are not necessarily in the core. These and other additional versions are discussed in Tijs and Driessen (1986).

4.2.3 Shapley Value

The Shapley value (Shapley 1953) is a solution concept that provides us with a unique solution to the cost allocation problem. The underlying idea is based on the assumption that the grand coalition is formed by entering the entities into this coalition one at a time. As each entity enters the coalition, it is allocated the marginal cost, and this means that its entry increases the total cost of the coalition it enters. The amount an entity receives by this scheme depends on the order in which the entities are entered. The Shapley value is just the average marginal cost of the entity, if the entities are entered in completely random order. The cost allocated to entity j is equal to

$$y_j = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (c(S \cup \{i\}) - c(S)) \tag{11}$$

Here, $|.|$ denotes the number of entities in the considered coalition. The quantity, $c(S \cup \{i\}) - c(S)$, is the amount by which the cost of coalition S increases when entity j joins it, here denoted by the marginal cost of entity j with respect to the coalition S. The Shapley value satisfies the efficiency property but does not necessarily satisfy the stability or the individual rationality properties.

4.2.4 Equal Profit Method

In many applications, the entities compute the relative savings and there is a desire to have an equal relative savings. One such approach, called Equal Profit Method (EPM), is suggested in Frisk et al. (2010). In this approach a Linear Programming (LP) model is solved where the model can be formulated as

$$\begin{aligned}
&\min f \\
&s.t. \\
&f \geq \frac{y_i}{c(\{i\})} - \frac{y_j}{c(\{j\})}, \forall i, j \\
&\sum_{j \in S} y_j \leq c(S), \forall S \subset N \\
&\sum_{j \in N} y_j = c(N) \\
&y_j \geq 0, \forall j
\end{aligned}$$

The first constraint set is to measure the pair-wise difference between the profits of the entities. The variable f is used in the objective to minimize the largest difference. The two other constraint sets define all stable allocations. In cases where the objective is not 0 (no difference between the entities), the reason is that there is a coalition that has an incentive to break out, i.e., the core constraints must be satisfied.

The EPM is related to a weighted version of the Constrained Egalitarian Allocation (CEA) method (Dutta and Ray 1991). The CEA method seeks to pick a point in the core that is as egalitarian as possible, i.e., where the allocated amounts are as equal as possible. We can also define a weighted version of the CEA method (Koster 1999). In order to relate the weighted CEA method to the method of Frisk et al. (2010), we set the weight of player i equal to $1/c(\{i\})$.

5 Technology

Building a coalition and managing it, require that quality information is shared between the collaborating entities. In all supply chains, information flows both upstream and downstream. For example, orders, sales forecasts, point of sales data or customer surveys, are sent top-down within the chain. In the opposite direction, information such as delivery plans, offers, catalogue, promotion, and availability (e.g., capacity or inventory) are sent from the suppliers to the customers.

The same principles also apply to financial flows. The payment will go from a customer to a supplier, whereas the credits, returns, or guarantees will go from the supplier to the customer. Financial flows involve financial institutions. All product and financial flows are supported by information flows. Typically, the flows link business entities two by two within the chain. However, collaborative logistics raise the need for higher connectivity as flows may be used by more entities.

This challenge has motivated great efforts to standardize the information flows. Different standards such as RosettaNet (www.rosettanet.org) provide shared guidelines in order to support timeless and effective interoperability within the supply chain. The aim is to eliminate the need for negotiating and agreeing on data definitions and formats with each trading entity, each time a possible transaction occurs. A common messaging interface enables the entities to exchange rapidly with many different entities by means of electronic data exchange technology and to reduce errors in data treatment and exchange.

Large corporations have started to use such standards to streamline their supply chain with their main customers (see PapiNet.org for examples). Although many standards exist, still many software enterprises develop their own model. In such cases, they typically use XML files to support the integration of the different technologies.

In logistics, application deals with different flows. Table 3 summarizes briefly the main information being exchanged when dealing with logistics.

Information can also be exchanged to inform on the status or characteristics of the main element of the value chain, as given in Table 4.

Table 3 Information exchange in logistics planning

Activity	Description
Plan	Includes a set of information defining when, where and how the different activities will be conducted within a business unit or a set of business units. We can differentiate the following plans: source, make, deliver, sales, and return
Order	Specifies a specific need from a specific business unit at a specific time. The orders can be production orders, transportation orders, sales orders, or return orders
Delivery	Specifies a delivery of a specific product or service at a specific business unit at a specific time. The deliveries can provide products/services to a business unit within the supply chain as well as to the end user or end customer. A delivery can be a return
Demand plan	A series of planned orders
Supply plan	A series of planned deliveries
Capacity plan	Defines the availability of the resources and their productivity
Forecast	Anticipation of an order, a delivery, or any plan
Execution parameters	A series of parameters defining how the execution of the different tasks should be conducted
Flow constraint	A series of constraints defining when and how the information flows can be exchanged
Execution updates	Information on how the execution really went

Table 4 Description of the main elements in the value chain

Element	Description
Product/service	Defines the characteristics of the product/service
Process	Defines the process in terms of input, resource consumption and outputs
Processor	Defines the characteristics of the processor
Inventory	Defines the number of products kept in stock
Entities	Defines a location

To support the information flows, platforms are needed. Building and sharing this type of infrastructure represent a challenging problem related to collaborative planning. Internet based technologies are providing many ways to connect the different entities and support their collaboration.

More specifically, agent-based technologies are rising as a new body of approaches building on distributed computing techniques. They intend to tackle the need for reactive, reliable, and (re)configurable operation management systems. An agent-based system may be defined as a system made of interdependent software agents designed to (a) individually handle a part of a problem such as planning an order or dispatching a task to a carrier and (b) collectively carry out specific higher functions such as planning a shared warehouse. Software agents generally exhibit characteristics that allow them to individually behave and interact with each other in order to fulfill the purpose of the entire system.

In such systems, information flows are supported by conversation protocols and messages. A conversation protocol links messages to form a conversation. As for

messages, they have their own purposes and move from one sender to one or many receivers. The platform to support the flows of messages can vary greatly going from a blackboard where messages are posted to a real collaborative logistics eHub.

Application of an agent-based platform for collaborative logistics is described in Audy et al. (2007b). It illustrates the Virtual Transportation Manager which is a web-based application permitting entities to post their transportation needs on a platform which thereafter optimizes the multiple pick-up and delivery transportation planning problem. The optimized routes, once accepted by the entities, are proposed to carriers.

6 Discussion

This paper sought to review some critical issues in building and planning a coalition with the aim of conducting collaborative logistics. As shown throughout the paper, the interest for this domain is rising both in the academic community as well as in industry. Even though new ideas and methods are provided to support the different decisions, many problems are still very difficult to deal with. These problems often call for interdisciplinary solutions. For example, in the process of building a coalition, some entities may be strong competitors. This type of relation has been called "co-opetition". In such a case, trust may play an important role in the decision process.

Legal issues are also very important. Many countries are concerned with potential collusive activities and therefore legislate to avoid them (e.g., antitrust law). Collaborative logistics projects need to address upfront the legislative limitation. Therefore, legal competencies may be needed. The sharing scheme may be analyzed in detail.

The sharing of the benefit in practice can be difficult, in particular when these benefits are intangible. For example, suppose there is a collaborative logistics project which permits an entity to access high value markets more easily, more rapidly and therefore develop them at low costs. What is the value of the increased geographical coverage or the faster deliveries?

Finally, the collaboration is rarely fixed in time. The environment changes constantly as well as the parameters considered when designing the collaboration. How should this dynamic be considered upfront? How often should the terms of the collaboration be reviewed?

References

Audy, J. F., D'Amours, S., & Rönnqvist, M. (2007a). Business models for collaborative planning in transportation: an application to wood products. In L. Camarinha-Matos, H. Afsarmanesh, P. Novais, & C. Analide (Eds.), *In IFIP international federation for information processing, establishing the foundation of collaborative networks* (Vol. 243, pp. 667–676). Boston: Springer. 10–12 septembre, Guimarães, Portugal.

Audy, J. F., D'Amours, S., & Rousseau, L. M. (2007b). Collaborative planning in a log truck pickup and delivery problem. *6th Triennial Symposium on Transportation Analysis* (p. 6), June 10–15, Phuket Island, Thailand.

Audy, J. F., D'Amours, S., & Rousseau, L. M. (2010). *Cost allocation in the establishment of a collaborative transportation agreement-an application in the furniture industry*. Journal of the Operational Research Society, (26 May 2010) doi:10.1057/jors.2010.53.

Beaudoin, D., LeBel, L., & Frayret, J. M. (2007). Tactical supply chain planning in the forest products industry through optimization and scenario-based analysis. *Canadian Journal of Forest Research, 37*(1), 128–140.

Cachon, G. P. (2003). Supply chain coordination with contracts. In *Handbooks in operations research and management science* (vol. 11, pp. 229–339). Elsevier.

Cederlund, J. P., Kohli, R., Sherer, S. A., & Yao, Y. (2007). *How Motorola put CPFR into action*. Supply chain management review, October 2007, 28–35.

Chen, F. (2003). Information sharing and supply chain coordination. In *Handbooks in operations research and management science* (vol. 11, 341–421). Elsevier.

D'Amours, F., D'Amours, S., & Frayret, J. M. (2004). *Collaboration et outils collaboratifs pour la PME Manufacturière*. Technical report, CEFRIO, http://www.cefrio.qc.ca/upload/1412_RapportfinalCollaborationetoutilscollaboratifs.pdf

Danese, P. (2006). The extended VMI for coordinating the whole supply network. *Journal of Manufacturing Technology Management*, 17(7), 888–907.

Dutta, B., & Ray, D. (1991). Constrained egalitarian allocations. *Games and Economic Behavior*, 3(4), 403–422.

Eriksson, J., & Rönnqvist, M. (2003). Decision support system/tools: Transportation and route planning: Åkarweb – a web based planning system. *Proceedings of the 2nd forest engineering conference, Växjö*, pp. 48–57, May 12–15.

Frayret, J. M., D'Amours, S., & Montreuil, B. (2004). Co-ordination and control in distributed and agent-based manufacturing systems. *Production Planning and Control, 15*(1), 1–13.

Frisk, M., Göthe-Lundgren, M., Jörnsten, K. & Rönnqvist, M. (2010) Cost allocation in collaborative forest transportation, *European Journal of Operational Research, 205*, 448–458.

Montreuil, B., Frayret, J. M., & D'Amours, S. (1999). A Strategic framework for networked manufacturing. *Computers in Industry*, 42(2–3), 299–317.

Moyaux, T., Chaib-draa, B., & D'Amours, S. (2007). The impact of information sharing on the efficiency of an ordering approach in reducing the bullwhip effect. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (SMC-C), vol. 37*(3).

Koster, M. (1999). *Weighted constrained egalitarianism in tu-games*, Center for Economic Research. Discussion Paper 107, Tilburg University.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in supply chain: the bullwhip effect. *Management Science, 43*(4), 546–558.

Lehoux, N., D'Amours, S., & Langevin, A. (2008) Dynamique des relations interentreprises: Mécanismes, barrières et cas pratique. *Revue Française de Gestion Industrielle 27*(4), 1–25.

Lehoux, N., D'Amours, S., & Langevin, A. (2009) Collaboration and decision models for a two-echelon supply chain: a case study in the pulp and paper industry. *Journal of Operations and Logistics, 2*(4), 1–17.

Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E. (1999). *Designing and Managing the Supply Chain: Concepts, Strategies, and Cases* (p. 321). McGraw-Hill/Irwin.

VICS, (2004). *CPFR Guidelines*. http://www.vics.org/home

Shapley, L. S. (1953). A value for n-person Games. *Annals of Mathematics Studies, 28*, 307–317.

Sprumont, Y. (1990). Population monotonic allocation schemes for cooperative games with transferable utility. *Games and Economic Behavior, 2*(4), 378–394.

Tijs, S. H., & Driessen, T. S. H. (1986). Game theory and cost allocation problems. *Management Science, 32*(8), 1015–1028.

Pilot Assignment to Ships in the Sea of Bothnia

Henrik Edwards

Abstract The Maritime Administration in Sweden initiated a study on the best allocation of pilots and its impact on costs and service levels. Pilots are required to guide ships from and to ports along the shipping fairways between the ports and open sea. Transports to and from rendezvous points at sea are made by small vessels. Normally the pilots are qualified to work on a subset of the shipping fairways. Average workloads are known from historic data, but the actual times for the required services are more or less known some 20+ h in advance. The problem studied was a historic data set covering 1 week of operations involving 66 pilot assignments and 21 pilots. A column generation approach was adapted. It includes the solving of shortest path problems in time–space graphs. Postprocessing is applied to handle legality constraints, which involve maximum work time shifts and minimum rest times in between. Set partitioning and set covering models were deployed to identify the final solution. Any multiple assignments associated with set covering formulations were handled by a myopic cost estimation procedure, and a small integer program for surplus removals.

1 Introduction

The Maritime Administration in Sweden was interested in a study on the best allocation of pilots and the impact on costs and service levels from system changes. In general the pilots are former sea captains that have chosen to change their line of work. Their work task is to guide ships from and to ports along the shipping fairways, stretching from the ports out to safe sailing water. Pilots board the ships at the start of the shipping fairway, in either of the both directions, and depart at the other end. Transports to and from points at sea are made by small vessels owned by the Maritime Administration. Normally the pilots are qualified to work on a

H. Edwards
Vectura Consulting AB, Box 46, 17111 Solna, Sweden
e-mail: henrik.edwards@vectura.se

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_23,
© Springer-Verlag Berlin Heidelberg 2010

subset of the shipping fairways. Average workloads are known from historic data, but the actual times for the required services are more or less known some 20+h in advance, when ships arriving to the southern Baltic provide arrival times to the shipping fairways. Departure times from the ports are also provided some 20–30 h in advance.

The pilot assignment problem displays many similarities with airline crew rostering problems that deal with assignment of airline pilots and cabin crew to anonymous tours of duties (using an expression from Ryan (2000)) Successful early implementation in Air New Zealand and development until 2000 is reported in, for example, Ryan (1992, 2000). Recent status of the work and methods used in this area, applied in some large international airlines, are reported in Kohl and Karish (2004).

Main differences are that the pilots, in general, work alone on the assignments, although boat transports are needed for carrying out the work. Boat transports occasionally turn out to be a scarce resource, so this complicating factor may optionally be included in future work. Also the demand for pilot assistance is subject to a more random variation from day to day, than the expected tours of duties for airline crew which is governed by time tables weeks ahead.

The problem studied was a historic data set covering 1 week of operations involving 66 pilot assignments and 21 pilots. A column generation approach was adapted. The column generation methodology has been widely used (see e.g., Jennergren and Dirickx, 1979) for a theoretical background. It includes the solving of SP (shortest path) problems in time–space graphs. Postprocessing, starting from the SP-solutions, is applied to handle legality constraints, which involve maximum work time shifts and minimum rest times in between. Better yet would have been to solve the column generation problem with a k-SP procedure, and keep the first legal solution or a set of legal solutions. Efficient k-SP algorithms for generating consecutive near-optimal shortest paths are described in Martins et al. (1998) and Martins and Santos (2000). The reasons for not doing this are primarily, a not fully operational code and secondly that the constructed networks result in many identical paths in terms of the important assignments, differing only in transport options. Generations of high quality work schedules to the master problem are, as in the airline crew rostering context, very important (see e.g., Kohl and Karish 2004).

The master problem is an linear programming (LP)-problem where duality variable values on assignment constraints are fed back to the SP subproblems. When the column generation was exhausted (no new columns were generated), the integer version of the problem formulated as a set partitioning problem was solved, although with some difficulty. Another option is to solve an (easier) set covering problem with a penalty, lower than using overtime, for multiple assignments. A postprocessing myopic cost estimation procedure was applied to determine the cost for giving the assignments to a single pilot, after which an integer program selected the optimal combination.

The SP subproblems were solved with a modified code from Gallo and Pallottino (1988). The linear programs and the integer programs were solved with LP_solve version 5.5 by Berkelaar et al. (2004).

This modeling effort is continued by the Maritime Administration, quoting (Maritime Administration, 2008): "To systematically enable the study and analysis of effects of differentiated service levels on the resource utilization and productivity of the pilot allocations a development work is ongoing, e.g. by cooperation with researchers at Campus Norrköping (part of Linköping University)."

2 Problem Description

The pilots are located along a number of pilot stations along the Swedish coast. Normally they are transported to the assignments on ships with car or boat, after which they guide the ship to port or from port to open water. The durations of the assignments are typically 1–2 h but may be longer. Each pilot has a qualification, a license stating the pilot area (the ship fairways), ship size and possibly type of ship determining what assignments he(she) can be assigned to. Some work time must be dedicated to training, in which case there is also an instructor present, that is, two pilots are working on the same assignment.

The service time per pilot and year is 155 days and nights, and of these 18 are unscheduled. The latter ones must be scheduled at the latest before start of the duty time period. Other work time is associated with overtime cost for the Maritime Administration. Normal duty hours are scheduled in advance.

The number of pilot assignments per year is approximately 40,000, and there are $230 \times 155 = 35{,}650$ days available. On the average this is short of 1.1 assignments per work day. Ships needing assistance notify the Maritime Administration at least 24 h beforehand. A definite request must be sent at least 5 h before start of the assignment.

Each duty day must include a rest period encompassing at least nine consecutive hours. Each rolling 24 h rolling horizon day must also contain a rest period with nine consecutive hours, which may be split. In the latter case the pilot receives the longer period. Shared rest must be agreed upon and cannot be ordered. The work shifts are restricted to at most 15 h/day. Further details are not handled in this project.

The task at hand is to carry out all the pilot assignments at the lowest possible cost while considering work time constraints (legality and scheduled duty days), qualification requirements, access to transport capacity (boats, see Fig. 1) and time constraints for the transports to/from the assignments and the pilot assignments themselves. Costs to consider are primarily variable costs for boats and taxi-transports, hotels, per diem and overtime costs. Taxi is the standard for land transports during duty time.

Strategic issues that may be the topic for studies with an efficient planning tool are

1. Is the pilot crew size appropriate?
2. Should efforts be made to enhance the flexibility by upgrading the pilot qualifications?
3. Should faster boat or helicopter transports be considered?

Fig. 1 Boats for transportation of pilots to and from boarding points

The number of pilot stations in Sweden is approximately 25, and there are some 50 public ports. On top of this there are some loading/unloading locations. Ship assignments can start at approximately 120 fairway sea starting points. Obviously we need to set up the transport time matrices for the relevant OD-relations with the different transport facilities.

3 Short-Term Planning

At some point it was decided to study an actual case using historical data covering 1 week in the Sea of Bothnia district. The model may of course cover many weeks. Input data required for the items in the problems [P1] and [P2] are

1. A number of pilot assignments are specified in time and space including qualification requirements. Alternative boarding points may exist (eg., if the ship is north bound or south bound). Start times and durations of the assignments along the shipping fairways are required. These data may be weather dependent, for example, in case of ice on the sea of Bothnia;
2. Pilots on duty and their qualifications. Possibly also remaining time until their qualification must be upgraded with the aid of instructor resources and training. Each duty day must comprise nine rest hours, with the option to split rest between

days. The duty day starts at a certain time, for example, 01:00, or when the first assignment for the day is commenced;
3. When required off-duty pilots can be called out for duty, in which case they work at overtime cost;
4. The start location of the first duty day, the home station or the alternative rest locations (hotels or other pilot stations) on multiple duty days. One constraint is that the pilot returns to his home station at the end of the duty period;
5. Coordinates (RT90) for
   - All start and end points for the shipping fairways where pilot assistance is required (including alternative starting points);
   - All pilot stations;
   - Home addresses;
6. A road transport network from the Sampers transport model, see Algers and Beser, (2000) for details, is used for determining car transport times between various destinations. The network is extended with links for boat and, possibly, helicopter transports.
   Private car is used between home and the home station, whereas taxi is used for other transports on land during duty time. Dedicated boats, including two boatmen, are used to reach boarding points at sea. One option could be to use a helicopter based at Midlanda airport (Sundsvall). In cases of using hotels or nonhome stations, this create time transports cost for coping with the rest periods;
7. The *value* of the assignments for the Maritime Administration. This value can be used for prioritizing between different assignments, for example, it could be based on a default value per dwt of the ships. The value should be large enough to make it worthwhile to carry out the assignment. When using overtime the value is not recognized. Initially it results in many assignments per pilot from the subproblems. A solution with zero value would not give any assignments, but only accumulate transport costs. In subsequent iterations the pricing mechanism, the shadow prices provide incentives to carry out (or skip) assignments;
8. In general, the use of overtime using a locally available pilot may result in a lower cost as compared to paying for long taxi transports and hotels. As already stated there are only 1.1 assignments per pilot and duty day, so a solution with increased use of overtime and fewer pilot duty days may, at least on the margin, be cost efficient for the Maritime Administration.

The problem described above is to minimize sum of cost for transports to/from/between pilot assignments and overtime costs while considering a number of side constraints. Here the columns $\{a_{ik}, b_{jk}\}$ involve that a work schedule k for pilot j must satisfy the work time rules regarding duty days, rest hours and overtime. Required qualifications must be satisfied and transport times between activities must be met. All of this is formulated as a set partitioning problem in (1)–(4), with relevant constraints implicitly satisfied by the columns used.

[P1] Set partioning

$$\min z = \sum_{i=1}^{M} c_i^o \cdot o_i + \sum_{j=1}^{N} c_j^u \cdot u_j + \sum_{k=1}^{K} c_k^a \cdot x_k \tag{1}$$

$$s.t. \qquad o_i + \sum_{k=1}^{K} a_{ik} \cdot x_k = 1 \quad for\, i = 1, 2, \ldots, M \tag{2}$$

$$u_j + \sum_{k=1}^{K} b_{jk} \cdot x_k = 1 \quad for\, j = 1, 2, \ldots, N \tag{3}$$

$$1 \geq o_i, u_j, x_k \geq 0 \tag{4}$$

where c_i^o is the cost associated with overtime usage for the assignments # i; c_i^u the costs/benefits for not using pilot # j for any assignment; c_k^a the sum of costs for work schedule # k; $a_{ik} = 1$ if work schedule # k includes assignment # i, 0 otherwise; $b_{jk} = 1$ if work schedule # k involves pilot # j, 0 otherwise; M is the number of assignment jobs; N the number of pilots; K the number of work schedule columns; $o_i = 1$ if assignment # i is carried out by using overtime, 0 otherwise; $u_j = 1$ if pilot # j has no scheduled assignments, 0 otherwise; $x_k = 1$ if work schedule # k is active, 0 otherwise.

The first set of constraints in (2) represents the demands for pilot assistance in time and space, and the remaining ones that each pilot either is assigned to one and only one work schedule or is inactive (so called convexity constraints). Should the demand not be satisfied with any of the active assignments, overtime capacity will be used.

The formulation in (1)–(4) is a set partitioning problem with equality constraints related to the demand. An alternative formulation is to use a set covering problem as in (5)–(8) with a penalty for multiple assignments. In general this problem, with a reasonable penalty, is easier to solve. Both options are used in the numerical experiments.

[P2] Set covering

$$\min z = \sum_{i=1}^{M} c_i^o \cdot o_i + \sum_{j=1}^{N} c_j^u \cdot u_j + \sum_{k=1}^{K} c_k^a \cdot x_k + \sum_{i=1}^{M} c_k^s \cdot s_i \tag{5}$$

$$s.t. \qquad o_i + \sum_{k=1}^{K} a_{ik} \cdot x_k - s_i = 1 \quad for\, i = 1, 2, \ldots, M \tag{6}$$

$$u_j + \sum_{k=1}^{K} b_{jk} \cdot x_k = 1 \quad for\, j = 1, 2, \ldots, N \tag{7}$$

$$1 \geq o_i, u_j, x_k \geq 0, s_i \geq 0 \tag{8}$$

where c_i^s is the cost penalty per unit over-coverage of assignment # i; s_i the over-coverage volume of assignment # i.

The discussed model is formulated as a deterministic problem with known demand for pilots during a full week. This will give an *over-optimistic estimation* of the possible cost savings, since demand in reality is uncertain. A more realistic evalution could include simulation of ship arrivals in time, space and numbers and running the allocation model on a rolling horizon basis. Such an elaborate evaluation was outside the scope of this project, and it would also require a model for the reference case, eg., the first-in-first-out (FIFO) rule applied in practice today.

3.1 Work Schedule Generation

With transport cost data according to items 5 and 6 above, time- and cost-matrices are determined between all points of interest, using the lowest cost for each mode of transport.

Now we construct a network model in time and space with connections to all ship assignment links (corresponding to the GAIA-schedule, produced by an existing administrative planning system). Data on ship fairways, arrival/departure times and possible alternative starting points are entered, as well as the value of the assignments. These links are connected to the rest of the network with optional transport links to/from pilot stations (or hotels). Also links from departure points at sea to arrival points at sea should be included, that is, the pilot may guide one vessel from the port out to sea, a transport boat goes out (with two boatmen) and picks up the pilot and awaits at sea an arriving vessel. The pilot boards the incoming vessel and guides it to port while the transport boat returns to its station. All transport links are provided time and cost values. An example is shown in Fig. 2 Off-duty links are also inserted to enable the search of a path from the start to the end of the planning horizon. Each one of these links corresponds by construction to *an offer you cannot refuse*.

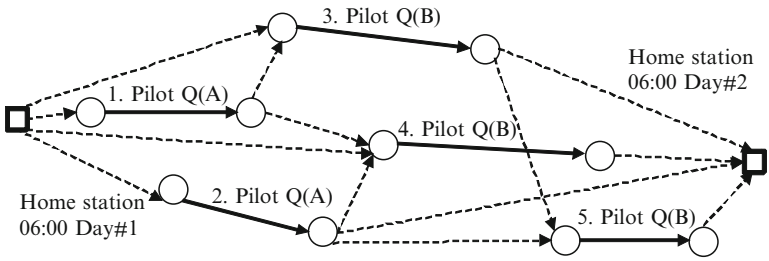


Fig. 2 Time–space network with five pilot assignments numbered 1–5, with qualifications A or B. Only a subset of possible links is shown. *Solid links* are the pilot assignments. Transport links are *dotted*

Shortest paths identifying the minimum *cost* in each network model are constructed, that is, we solve

$$\begin{aligned} &\textit{min Cost} = -\sum(\textit{Assignment values} + \textit{Shadow prices}) + \sum \textit{Transport costs} \\ &\textit{subject to network flow constraints} \end{aligned} \tag{9}$$

Unfortunately, there are side constraints, legality conditions associated with work schedules that must be satisfied: minimum nine consecutive hours rest time, and maximum 15h of duty time, per day. These cannot be included directly in the SP-algorithm. State variables can be set up at the nodes to reflect the accumulated work and rest times, and a method based on a topologically sorted graph can be used to ascertain that only one search of each node will be needed. This will not be enough since we cannot know what combination of objective value and state variable values is the best one. Ideally a k-SP algorithm would be used and the best legal solution(s) could be inserted into the master problem. While not having operational software for this, we formulated a dynamic program that kept an optimal subset of the SP-schedule while guaranteeing a legal solution.

In the example in Fig. 2 we find that the five pilot assignments can be carried out with two pilots, each of which having the qualification AB (both A and B), namely with the schedules

$1 \rightarrow 3 \rightarrow 5$

$2 \rightarrow 4$ (assumes a helicopter transport between the jobs)

It is assumed that both these schedules are compatible with a rest period of 9 h. If not, more pilots need to be engaged. This is also the case if any of the pilots does not have qualification AB. Should helicopter transport not be allowed it may be better to combine assignments 2 and 5.

The operative task at hand is to distribute the pilot assignments daily and utilize the pilots on duty in the best possible manner. One common way to assign pilots is to use a cyclical, FIFO procedure, in which case assignments are given to the pilot next in the queue. The idea behind this is to achieve a fair distribution of the assignments which is preferred among the pilots because they are awarded extra payments for each guided ship.

The work schedule should be constructed so that the expected yearly demand for pilots can be handled. A good basis for this is the historical data, but also information from shippers on future required services. Some additional capacity should be available to cope with unexpected demand peaks, illnesses etc.

3.1.1 Network Flow Model Details

As illustrated in Fig. 2 our network model consists of links for pilot assignments and transport links, both for physical transports but also waiting times at stations and hotels. Nodes represent locations in time and space. The time scale is in minutes and the geographic locations are represented by unique numbers in different intervals ($1{,}000 + k$ for pilot station number k, $2{,}000 + m$ for port number m, and $4{,}000 + n$ for boarding point number n)

All pilots in the system must pass through the network model from time 0 to the end of the planning horizon 10080(= 7*1440 min). Nonworking days are represented by pilot specific sets of compulsory day-off links. During duty days there are optional assignments and transports to and from jobs. Rest time and other time are spent either at a pilot station, a hotel (when ports are located far away from pilot stations) or in ports waiting for the next assignment.

The side constraints are handled in the following manner.

1. Solve SP. Check legality: cumulative work time must not exceed 15 h. Between work shifts a minimum rest period of 9 h is required. If OK, use the solution, otherwise go to step 2.
2. We commence with the solution from Step 1. This postprocess is formulated as a DP (dynamic programming) problem, where the decisions concern which of the SP assignments to keep. The state variables are the current cumulative work time and off-work time, respectively; the decision for each assignment is whether or not to keep it, and the objective is the same cost minimization as in the SP subproblems.

A number of assignments are removed from the SP-solution according to the DP-solution to make it legal. The objective function of the modified solution is recalculated.

A small modification is that transports not made in association with a job assignment are penalized with a small cost (1 SEK) to avoid interrupting a period of rest with unnecessary work.

3.2 The Master Problem

A column generation method is suggested to solve problem [P1] or [P2], in which the columns are generated with the work schedule generator from Sect. 3.1. The generated schedules are weighted together in the master program formulated as an LP-problem that combines the schedules in the best possible manner. The master is formulated as a set partitioning problem implying that each ship assignment should be assigned to at most one pilot. Convexity constraints requiring that the pilots may be used for one work schedule only, or none at all, are inserted. One overtime schedule with an anonymous pilot is inserted for each ship assignment to ascertain that a feasible LP-solution always can be identified.

In normal cases the master problem does not identify a feasible, integer solution, but dual variable values (shadow prices) reflecting the marginal effects on the objective function of the various constraints are obtained. This information is fed back to the subproblems, where they are used to reevaluate the assignments inside the network models. New work schedules are identified and the master problem is solved again. This process is repeated until no new work schedules are generated (or another stopping criterion is satisfied).

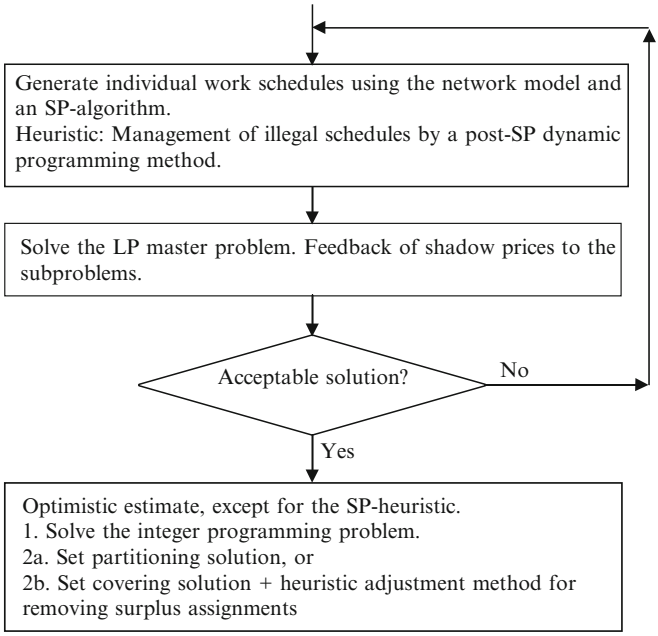


Fig. 3 Flow chart for a column generation approach to the pilot assignment problem

When the column search is exhausted the final version of the master can be solved with an MIP (mixed integer program) solver according to formulation [P1] or [P2]. The set covering solution from [P2] can be shrunk into a feasible solution by removing multiple assignments from all of the used work schedules but one. This shrinking is carried out by a myopic, heuristic, costing procedure involving computing the column costs without the multiplying covered assignments one by one. Then a small integer program assigned these at the lowest cost (=largest saving). The methodology is illustrated in Fig. 3.

4 Input Data Example from the Sea of Bothnia

In this study we work with a data set from week 41, 2004. The coast line covered stretches almost 300 km from Gävle to Härnösand. The pilot stations are located close to Gävle and Sundsvall with two additional ones in between.

Estimated distances with private car (taxi) between different places are based on network data from the Sampers-model, see Edwards, (2006) for details. This is the case also for other detailed input data that does not fit in here. A summary of the data is given in Table 1.

Estimated variable cost for the pilot boats (with march speed 9 kn) is based on diesel consumption according to the Marine Administration at approximately 80 L

Table 1 Summary of input data

Description	Count	Transport	Speed range (kn)	Cost (SEK/h)
Pilot stations	4	2 boats/stn	9–35	370–920
Airport	1	1 helicopter	150	6,000
Ports	25			
Boarding points: sea	7	Associated with their closest pilot station		
Pilots	21			
Qualifications		5–18 ship fairways, median = 5		
Assignments	66	Duration = 20–355 min, average = 87 min, median = 75 min		
Transport times		50% < 60 min, max 3:30 h		
Qualifications		5–18 ship fairways, median = 5		

of diesel oil per hour. The pilot boats with a march speed at 23 and 35 kn, respectively consume approximately 200 L/h. No taxes are paid for the fuels (none of the energy, CO_2 or VAT tax components). Each liter of diesel oil then costs 4.60 SEK with *a land price* at 10 SEK ($= 10/1.25 - 3.34$). In general, the speeds of the boats are more important than the costs, since the number of man-hours is a more important factor. The accessibility of a boat transport is assumed given in our study, but conflicts may arise in the solution due to too many assignments in a too short time. This is also the case should the fastest boat be double booked.

4.1 Values, Costs, and Other Control Parameters

We are looking for the resource utilization given the demand for ship assignments. Revenues are the fairway dues and the pilot fees that must be paid for the ships to the Maritime Administration. The contributions to the objective in the assignment problem are primarily awards used for pilot assignments (a control parameter to stimulate work activities), and possibly, estimated values of unused duty days. The cost components concern overtime, transports and hotels. The data in Table 2 has been used in a first comparison of the model performance compared to the actual assignments in week 41, 2004.

4.2 Reference Case Value

The historic work schedule is valued at −96,974 SEK. In summary, 5 of the 66 assignments are performed outside regular duty periods. There are 54 active and 6 passive duty days. There is one boat transport conflict (and at least one case with a boat transport between ships at the sea boarding point). It should be possible to formulate a model that also deals with the boat transport capacities.

Table 2 General parameters in our pilot assignment study

Description	Value/cost
Cost for private car use	2.50 (SEK/km)
Average cost per km for taxi	20 (SEK/km)
Average hotel cost	800 (SEK/day)
Cost for over-time assignment as a share of the assignment award. 0 corresponds to zero extra cost	0 (%)
Value of an unused duty day. It may be valued at 8 h/day ×200 SEK/h × 1.6 (LKP) = 2,560 SEK/day. Default: 2 SEK	2 (SEK/day)
Cost per over-time hour with 60% LKP (social costs). Minimum time is 3 h including transport to and from home	640 (SEK/h)
Time to pick up and drop a pilot when using helicopter	10 (min)
Time of day when assignments may commence in a new duty period	60 (min)
Value of carrying out an assignment. This is subtracted from the final objective	1,000 (SEK/asg)
Default car distance when calling in over-time resources. Average distance between the pilot's home and the station	100 (km)
Forbid or allow delays of the assignment start. Any allowed delays are assignment specific	0 (boolean variable 0/1)
Penalty for over-covering assignments in the master (set to 10,000 for the LP-solutions). Set to > 1,000 to get the set partitioning solution	50 (SEK/asg)

5 Column Generation Solutions

We have solved the problem described with the column generation method from Sect. 3. A number of scenario results are presented with variations of MIP formulations, assignment values, and increased flexibility in terms of available pilots. For an earlier scenario with possibilities to delay assignments we ran probably into a book keeping problem with the links.

(1 & 2) Optimal Solution compared to the reference case (w 41, 2004).

To motivate carrying out assignments in the subproblems a control parameter award at 1,000 SEK per assignment has been used. This is subtracted from the final solution; otherwise the objective value would increase by 66,000 SEK, provided all were performed at regular duty hours. When the column generation was exhausted, the final integer problem was solved both according to [P1] and [P2], including a heuristic for the surplus allocations in the latter case. There were approximately 30 over-coverings. The results of the solutions are summarized in Table 3.

(3) A solution with full qualifications for all pilots.

In this scenario all pilots can operate on all the shipping fairways. Obviously this is a relaxation of the problem since more opportunities are opened up. The solution obtained with the set coverage variant is marginally better than its counterpart above.

Table 3 Summary of the reference case and the scenario solutions. PART corresponds to a set partitioning solution, and COVER to a set covering solution including the myopic heuristic

	Ref w 41, 2004 (Sect. 4.2)	(1) Opt soln PART	(2) Opt soln COVER	(3) Full qual for all pilots COVER	(4) Extra work day available COVER	(5) Opt soln PART (no asg award)
Optimistic estimate	–	−57,511	−57,511	−55,844	−56,679	−58,263
Best found solution	−96,974	−61,011	−62,446	−61,612	−56,668	−61,907
# Asg on overtime	5	0	0			
# Pilots used (14 on duty)	13	13	14	14	14 +1 extra	13
# Utilized duty days	54	55	60	60	61	55
# Unused duty days	6	5	0	0	0	5
MIP CPU-time		4 h	≈1 min	≈1 min	≈1 min	≈10 min

(4) A solution with flexible allocation of regular duty time.

This scenario is the same as (2) with the option to hire an extra duty day at any pilot station at a cost of 2,540 SEK/day. One extra day was acquired.

(5) Optimal solution without an assignment award.

This scenario is the same as (1) but without the assignment award.

5.1 Some Solution Details

The results are summarized in Table 3. The optimal, or near-optimal, solutions have objective values in the range 57,000–62,000 SEK as compared to the actual solution valued at 97,000 SEK. The optimistic estimates are only somewhat lower, so the duality gaps are rather small. A full implementation of the method ought to be valuable. However, there are a number of unrecognized complicating factors, for example,

1. The ship assignment demand is unknown, even during a 1 week term.
2. There may be unofficial constraints on the free allocations of pilots.
3. The pilot boat capacity has not been considered.

When using overtime we have assumed a taxi cost for 100 km for the one-way trip between the home and the station. This fact may over-emphasize the importance of avoiding overtime, since each instance of overtime incurs taxi costs at 4,000 SEK.

Should the distance only be 50 km on the average, the over-time costs in the reference case would shrink by $5 \times 2{,}000$ SEK which is almost onethird of the savings. With other transport costs it may very well be worthwhile to use overtime for a pilot in the neighborhood.

6 Conclusions

A column generation method has been formulated and implemented to solve a pilot assignment problem in a case study for the sea of Bothnia. With the used valuations, and an artificial fixed demand case, it was possible to generate considerable savings. It should be possible to use the method for routine pilot allocation operations, but there are many more options available, of which some have been tested, namely:

1. Enable the possibility to analyze legality changes and their implications.
2. Study the effects of improving the flexibility of the organization (qualifications, hiring, delaying services, etc.).
3. Enable the study of different means of transports (helicopter has been discussed).
4. Compare impacts of using various pilot staff levels and free-lancing pilots.
5. Estimate the demand for pilots on a long-term basis given ship assignment forecasts.
6. Location of pilot stations.

On the matter of the solution methods we have used the freely available LP_solve by Berkelaar et al. (2004). Could the matter of the many practically identical shortest paths be managed, the shortest path problems with legality side constraints should preferably be solved with a k-SP algorithm. From the k-SP solutions, the first, or a set of, legal columns should be inserted into the master problem. With the currently used heuristic approaches for solving the subproblems and the myopic assignment heuristic applied after the set covering MIP, we cannot guarantee that the optimal solutions have been identified. On the other hand, given the many other complicating issues, the current model performs very well.

6.1 Future Work Suggestions

There are many ways to continue with this work of which a few are listed below. As mentioned in Sect. 1, some of these aspects are included in a continuation project by the Maritime Administration.

- Analyze the resource utilizations in terms of manpower and transport facilities.
- Consider pilot assignments in conjunction with suggested changes of work time rules.
- Consider effects of pilot qualification enhancements.
- Utilize models for handling uncertainty in pilot service demand.
- Include possibility to schedule instruction and qualification maintenance.

Acknowledgment The author wishes to thank the Maritime Administration for the opportunity to carry out this interesting and challenging study.

References

Algers, S., & Beser, M. (2000). *SAMPERS – The New Swedish national travel demand forecasting tool*. Sweden: Transek, Solna

Berkelaar, M., Eikland, K., & Notebaert, P. (2004). LP_solve 5.5, Open source (Mixed-Integer) Linear Programming system, GNU LGPL (Lesser General Public License)

Edwards, H. (2006). Lotsallokering i Bottenhavets sjötrafikområde. Project report in Swedish for the Maritime Administration

Gallo, G., & Pallottino, S. (1988). Shortest path algorithms. *Annals of Operations Research, 13*(7), 3–79

Jennergren, L. P., & Dirickx Y. M. I. (1979). *Systems analysis by multilevel methods: With applications to economics and management.* Chichester: John Wiley and Sons

Kohl, N., & Karish, S. E. (2004). Airline crew rostering: Problem types, modeling, and optimization. *Annals of Operations Research, 127*(1–4), 223–257

Maritime Administration (2008). Three-year plan for 2009–2011, Norrköping

Martins, E., Pascoal, M., & Santos, J. (1998). The k-shortest paths problem Department of Mathematics, University of Coimbra, Portugal

Martins, E., & Santos, J. (2000). A new shortest paths ranking algorithm. *Investigação Operacional, 20*(1), 47–62

Ryan, D. M. (1992). The solution of massive generalised set partitioning problems in aircrew rostering. *The Journal of the Operational Research Society, 43*(5), 459–467

Ryan, D. M. (2000). Optimization earns its wings. resource document. *OR/MS Today, April* http://www.lionhrtpub.com/orms/orms-4–00/ryan.html. Accessed 23 March 2009

Transportation Planning and Inventory Management in the LNG Supply Chain

Henrik Andersson, Marielle Christiansen, and Kjetil Fagerholt

Abstract In this chapter, the LNG supply chain is introduced, and two planning problems related to the transportation planning and inventory management within the chain are presented. The two problems have different characteristics and reflect the planning situations for a producer and a vertically integrated company, respectively. Both problems are formulated as mixed integer programs, and possible solution methods are briefly discussed.

1 Introduction

The energy demand has been constantly increasing during the past decades. While other essential energy sources, such as oil and coal, have become less attractive due to increased prices and environmental considerations, natural gas has become more important to fulfill this demand. Estimates predict that natural gas will constitute 28% of the total energy consumption in 2020, compared with 23% in 1999 (EnergyInfo 2007).

Over shorter distances, natural gas is usually transported in pipelines. However, this is not an economically viable alternative for longer distances, especially across the seas. Technological advances in the maritime industry have made ships a good alternative for long distance transportation of natural gas. This can be done by cooling down the gas at atmospheric pressure to a temperature of $-162°C$ ($-260°F$). At this temperature the natural gas reaches its liquid state and turns into Liquefied Natural Gas (LNG). This also reduces the volume of the gas by a factor of 610 (EIA 2003), which makes transportation and storage more efficient.

While the capacity of the world fleet of LNG ships was only about 5 million m^3 in 1980, it had increased to about 35 million m^3 in 2007 and is expected to reach about 55 million m^3 by 2010 (NGR 2007). The average capacity of the ships

H. Andersson (✉), M. Christiansen, and K. Fagerholt
Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Gløshaugen, Alfred Getz vei 3, 7491 Trondheim, Norway
e-mail: henrik.andersson@iot.ntnu.no; mc@iot.ntnu.no; kjetil.fagerholt@iot.ntnu.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_24,
© Springer-Verlag Berlin Heidelberg 2010

has also grown, from 140,000 m^3 in 2005 to 180,000 m^3 in 2008 (NGR 2007). The consumption of LNG will increase in a similar way. It is expected that LNG will represent around 20% of the OECD gas supply already in 2010, and 14–16% of global gas demand by 2015 (NGR 2007). The number of supply sources and demand regions has also increased.

Managing the LNG supply chain is an important task. The LNG supply chain can be defined as all processes from extraction of the natural gas until it is consumed by the end users. In this chapter, we focus on the shipping decision making. However, since shipping decisions can be closely related to inventory management decisions at the liquefaction plants and regasification terminals, we also include these processes of the supply chain. This results in combined transportation planning and inventory management problems, which are often referred to as inventory ship routing problems. Christiansen et al. (2004) present a review of ship routing problems in general, while Christiansen et al. (2007) present mathematical models for some inventory ship routing problems.

Knowing that the sales price of one typical LNG shipload can be 20–35 million USD and that the daily time charter rate for a conventional LNG ship is about 70,000 USD, there is a lot at stake. The LNG supply chain has become more complex, partly due to the increased volumes and higher number of supply sources and demand regions. Traditionally, most LNG ships have been tied up to specific contracts and trades, where they would be dedicated to shuttle between given liquefaction plants and regasification terminals. This is about to change. The rigidness of contracts seems to loosen up, and there is also a growing spot market for LNG, which creates more business opportunities serving different markets and introduces more flexibility in utilizing the LNG ships. Moreover, this also contributes to making the supply chain planning more challenging.

Because of this increased complexity and importance we have seen a significant research interest in optimizing the LNG supply chain for the last years (see, for instance, Grønhaug and Christiansen 2009 and Stremersch et al. 2008). The aim of this chapter is to contribute to this research by presenting two typical combined transportation planning and inventory management problems for the LNG business. A mixed integer programming (MIP) model will also be given for each of the two planning problems. However, we will not focus on solution approaches and computational results for the problems. Both solution approaches and computational results can be found in Grønhaug and Christiansen (2009), Grønhaug et al. (2010), and Moe et al. (2008) for real world problems.

In Sect. 2, we describe the LNG supply chain in more detail. Then, we present some common problem characteristics and notation in Sect. 3, while Sects. 4 and 5 describe the two LNG supply chain planning problems with their corresponding MIP models. In Sect. 4, we consider a problem for an LNG producer with only one production facility. The producer is in charge of the inventory at the liquefaction plant as well as ship transportation of the LNG to its customers. We do not focus on decisions related to sales, and assume that the demand is given as input to the model. In Sect. 5, we consider a vertically integrated company, which controls several liquefaction plants, as well as being in charge of the inventories at

the regasification terminals. Here, we also include sales decisions into the model. For both problems presented in Sects. 4 and 5 we assume that the production volumes are given. Solution methods relevant for both problems are briefly presented in Sect. 6. Finally, some concluding remarks are given in Sect. 7.

2 The LNG Supply Chain

The LNG supply chain, illustrated in Fig. 1, begins with the natural gas being extracted and sent through pipelines to nearby liquefaction plants. Impurities are removed from the gas and it is cooled down to its liquid state and stored in tanks built with full-containment walls and systems to keep the gas liquefied. The LNG is then transported by ships to its destination terminal. There, it is again stored until it is regasified and finally sent through pipelines to the end users, which can be power plants, industrial customers, and households.

In Fig. 1, the parts of the supply chain we focus on in this chapter have been marked by a rectangle. Here, the main goal is to design an optimal plan to transport the LNG from the liquefaction plant(s) to the regasification terminals. However, the transportation decisions must be integrated with the inventory decisions to ensure that no inventory level exceeds its upper limit or falls below its lower limit. If an inventory level at a liquefaction plant exceeds its limit, the producer must shut down production, which will have major consequences. If an inventory level at a regasification terminal falls below its limit, it may result in lost sales and/or contract breach with its customers.

It is common to distinguish between three planning levels with different time horizons when planning the supply chain (Stremersch et al. 2008).

Long-term (strategic) planning: Long-term planning typically includes decisions about investments and long-term contracts that will have an impact many years ahead.

Annual delivery programme (ADP) set-up: This is more of a tactical planning problem with a typical planning horizon of 12–18 months. When setting-up the ADP, the aim is to determine an optimal fleet schedule, including the delivery dates (or time windows) at the different customers' terminals. This fleet schedule must also satisfy inventory constraints, as well as customer contract constraints.

Operational planning: This deals with updating fleet schedules due to various logistics, economic, or contractual reasons. Examples of logistics reasons can be rescheduling due to unplanned events, such as equipment breakdown or ship delays.

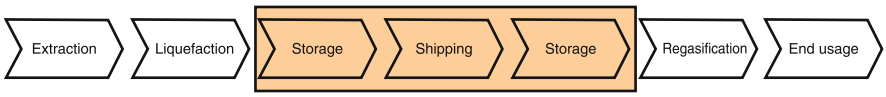


Fig. 1 The LNG supply chain

An example of an economic reason is when spot market prices change; new sales or purchase opportunities may create needs for rescheduling. The typical length of the operational planning horizon is 3 months.

In this chapter, we focus on the tactical and operational planning levels and the models presented in Sects. 3 and 4 can be applicable to both.

3 Common Problem Characteristics and Notation

There exist some common characteristics and notation for the problems presented in Sects. 4 and 5, which will be described in the following. Much of this material can be found in Grønhaug and Christiansen (2009) and Grønhaug et al. (2010).

Time periods: We assume that the planning horizon is discretized in time periods. The length of a time period is one day, since a port visit typically takes one day. The sailing times between ports are also accurately enough if given in a number of days.

Inventory management at a liquefaction plant: A particular company may produce several qualities of LNG, like lean and rich LNG. Here, LNG is considered as a homogeneous product. We assume that the production volumes of LNG at a liquefaction plant are given for each time period of the planning horizon. The inventory capacity at a plant is also given.

Berth capacities: The ports have limited capacity, such that a given maximum number of ships can visit each port in each time period. However, it is possible to wait outside a port before loading and unloading.

Sailing and the fleet of LNG ships: The fleet of ships is heterogeneous and assumed to be fixed for the planning horizon we consider. The ships may have different cost structures and load capacities.

The cargo hold is separated into several tanks. Here, we assume that all ships load and unload a full shipload. This means that there are no successive calls to ports of the same type (liquefaction or regasification). Full shiploading and unloading happen most often in practice as port tolls are high. See Grønhaug and Christiansen (2009) for models where successive port calls to regasification terminals are considered.

Some of the LNG is vaporized during a voyage and this gas, called boil-off, is used as fuel. Some quantity, given by a constant rate of the cargo capacity, is boiling off each day during a voyage. In order to keep the cargo tank cool, some LNG must always be left in the tanks except just before loading.

Figure 2 shows four snapshots of a voyage for a ship containing two tanks. In Fig. 2a, the ship leaves the liquefaction plant fully loaded and the storage there is empty. The ship sails to a regasification terminal, and it has to visit this terminal before the storage is empty. In Fig. 2b, we see that some of the gas has evaporated while sailing. The ship can then unload the rest of the LNG, except the LNG that is boiling off on the leg from the regasification terminal to the liquefaction plant; see Fig. 2c. Finally, in Fig. 2d, the ship returns to the same liquefaction plant and the

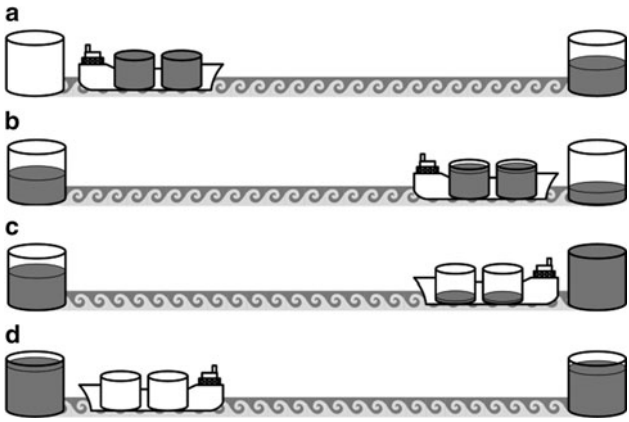


Fig. 2 Ship inventory

ship is just empty when it arrives at the port. However, the ship could as well have sailed to another liquefaction plant.

During the planning horizon each ship sails a geographical route. This route together with arrival times at each port visit makes the schedule for each of the ships. The ships are always fully loaded at the liquefaction plants and the unloaded quantity is adjusted for the boil-off. In principle, it is possible to generate all schedules with associated unloaded quantities in advance, but for real instances the number of schedules might be very high.

There is no natural depot for the ships. The initial position of a ship may be at a port or a point at sea. Thus, the ships might be empty or loaded, and there is a set of first port call candidates in its route. Furthermore, there is no requirement for a specified position for any ship at the end of the planning horizon.

Cost structure: Since we do not consider the option to change the fleet size during the planning horizon, fixed costs for the fleet are disregarded. The variable sailing costs consist of fuel and diesel oil costs. In addition, the LNG ships are charged port tolls when visiting ports and these costs depend on the size of the ship.

Common notation: In the mathematical description of the problem, each port is represented by an index i, and the sets of liquefaction plants (pickup ports) and regasification terminals (delivery ports) are given by $\mathcal{N}^{\mathcal{P}}$ and $\mathcal{N}^{\mathcal{D}}$, respectively. Then, $\mathcal{N} = \mathcal{N}^{\mathcal{P}} \cup \mathcal{N}^{\mathcal{D}}$ is the set of all ports. The set of time periods (days) in the planning horizon is given by $\mathcal{T}$. Furthermore, $\mathcal{V}$ is the set of available ships.

There are some ports where inventory management must be considered as part of the problem. The inventory levels at these ports must be within given lower and upper limits, $[S_i^-, S_i^+]$, during the planning horizon. The production volume at port i in time period t is given by P_{it}. A ship is always fully loaded at liquefaction plants and the capacity of ship v is V_v.

The continuous variable s_{it} represents the inventory level at port i in time period t. Note that s_{i0} is a parameter representing the initial inventory at the beginning of the planning horizon.

B_i denotes the maximum number of ships that can be loaded or unloaded at port i in a given time period.

4 A Planning Problem for a Producer

In this problem, the LNG supply chain is studied from a producer's perspective. The producer controls one liquefaction plant and serves a number of regasification terminals using a fleet of ships. Since there is only one liquefaction plant and full shiploads are assumed, the challenge is to sequence and schedule voyages and to assign them to ships.

The LNG is typically sold on long term contracts, which specify the quantity of gas that should be delivered each year. The annual quantity is often given by an interval, giving the producer some volume flexibility. The contracts also state that the LNG deliveries should be evenly distributed over the year, without specifying the exact dates and quantities. The evenly spread clause can be modeled in many different ways. Usually, the annual quantity demanded by a customer is split into monthly demands, reflecting both the evenly spread clause and seasonal variations. Then, the total delivered quantity each month is measured against the monthly demands and failure to meet the demand is penalized.

The problem is to design a minimum cost plan that assigns a number of voyages to each ship. In addition, the inventory level at the liquefaction plant must always be between the specified upper and lower limits. Considerations must also be taken to the berth capacity at the liquefaction plant and the contractual agreements.

Instead of introducing schedules consisting of geographical routes and arrival times for the entire planning horizon, we operate with voyages in the underlying model of this problem. Normally, there will be a number of voyages in each schedule for a particular ship. In Fig. 2, we see a voyage starting in a liquefaction plant, then visiting a regasification terminal before returning to the liquefaction plant.

In addition to the common notation presented in Sect. 3, the following notation is needed. A voyage r includes loading at the liquefaction plant, the sailing to and unloading at a specific regasification terminal, the sailing back to the liquefaction plant and the waiting until a new voyage can begin. All voyages have a specified starting time; hence, each physical voyage is duplicated throughout the planning horizon using index t. Let $\mathcal{R}_v$ be the set of all voyages that can be assigned to ship v. Since the destination and the time to the next loading is specified in the voyage and just one regasification terminal is visited on each voyage, the quantity delivered to the regasification terminal by ship v on voyage r starting at time period t, Q_{vrt}, as well as the cost of ship v operating voyage r starting at time period t, C_{vrt}, can be calculated. Due to, for example, different seasonal

weather conditions over the planning horizon, both cost and delivered quantity can be time dependent. The delivered quantity is the capacity of the ship minus boil-off; this means that the longer the voyage the less is delivered. The set of time periods (days) in the planning horizon is aggregated into months, m, which are used to model the evenly spread clause. The set of all months is $\mathcal{M}$. The costs associated with over- and under-delivery to regasification terminal i in month m is denoted by C^O_{im} and C^U_{im} and the corresponding demand is denoted by D_{im}. The total delivered quantity at regasification terminal i should be within given lower and upper limits, $[H_i{}^-, H_i{}^+]$. There are berth capacity constraints at the liquefaction plant, while it is assumed that no berth constraints are needed at the regasification terminals.

To make the model more readable, the following indicators are introduced: A^C_{vrtim} is 1 if voyage r of ship v starting at time period t delivers to regasification terminal i in month m, and 0 otherwise; $A^R_{vrt\tau}$ is 1 if voyage r starting at time period t is operated by ship v at time period τ, and 0 otherwise.

The variables in the model are δ_{vrt} which is 1 if voyage r starting at time period t is assigned to ship v, and 0 otherwise. o_{im} and u_{im} are contract deviation variables and represent over- and under-delivery at regasification terminal i in month m, respectively. To simplify the presentation, the variable h_i, representing the total delivered quantity to regasification plant i is introduced. Since there is only one liquefaction plant in the problem considered here, the variables and parameters representing the berth capacity, the inventory level and limits, and the production volume at the liquefaction plant introduced in Sect. 3 are used without indices for the liquefaction plant. With this notation, the following model can be formulated:

$$\min \sum_{v\in\mathcal{V}} \sum_{r\in\mathcal{R}_v} \sum_{t\in\mathcal{T}} C_{vrt}\delta_{vrt} + \sum_{i\in\mathcal{N}^D} \sum_{m\in\mathcal{M}} \left(C^O_{im} o_{im} + C^U_{im} u_{im}\right), \tag{4.1}$$

subject to

$$s_t - s_{t-1} + \sum_{v\in\mathcal{V}} \sum_{r\in\mathcal{R}_v} V_v \delta_{vrt} = P_t, \quad \forall t \in \mathcal{T}, \tag{4.2}$$

$$\sum_{v\in\mathcal{V}} \sum_{r\in\mathcal{R}_v} \sum_{t\in\mathcal{T}} Q_{vrt} A^C_{vrtim} \delta_{vrt} + u_{im} - o_{im} = D_{im}, \quad \forall i \in \mathcal{N}^D, m \in \mathcal{M}, \tag{4.3}$$

$$\sum_{v\in\mathcal{V}} \sum_{r\in\mathcal{R}_v} \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{M}} Q_{vrt} A^C_{vrtim} \delta_{vrt} - h_i = 0, \quad \forall i \in \mathcal{N}^D, \tag{4.4}$$

$$\sum_{v\in\mathcal{V}} \sum_{r\in\mathcal{R}_v} \delta_{vrt} \le B, \quad \forall t \in \mathcal{T}, \tag{4.5}$$

$$\sum_{r\in\mathcal{R}_v} \sum_{t\in\mathcal{T}} A^R_{vrt\tau} \delta_{vrt} \le 1, \quad \forall v \in \mathcal{V}, \tau \in \mathcal{T}, \tag{4.6}$$

$$H_i^- \le h_i \le H_i^+, \quad \forall i \in \mathcal{N}^D, \tag{4.7}$$

$$S^- \le s_t \le S^+, \quad \forall t \in \mathcal{T}, \tag{4.8}$$

$$u_{im}, o_{im} \geq 0, \quad \forall i \in \mathcal{N}^D, m \in \mathcal{M}, \tag{4.9}$$

$$\delta_{vrt} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v, t \in \mathcal{T}. \tag{4.10}$$

The objective function (4.1) consists of the cost of operating the voyages and the cost of over- and under-deliveries. The inventory balances for the liquefaction plant is given in (4.2). In constraints (4.3), the quantity delivered each month is measured against the demand, and o_{im} or u_{im} becomes positive if too much or too little is delivered. The total quantity delivered to each regasification plant is specified in constraints (4.4). Constraints (4.5) control the number of ships that can load in the same time period. The scheduling constraints (4.6) state that each ship can only operate one voyage each time period, and constraints (4.7)–(4.10) give upper and lower bounds on the total delivered quantities and inventory level, non-negativity on the contract deviation variables, and binary requirements on the voyage variables, respectively.

5 A Planning Problem for a Vertically Integrated Company

We assume that the vertically integrated company is responsible for the inventory management at all liquefaction plants and regasification terminals in addition to the transportation between these plants and the sales at the regasification terminals. Figure 3 shows an example of a routing plan for a problem consisting of three liquefaction plants, four regasification terminals, and two ships. Some ports are called several times during the planning horizon by the same or different ships. Liquefaction plant 1, P1, is just called once, while P2 is called three times during the

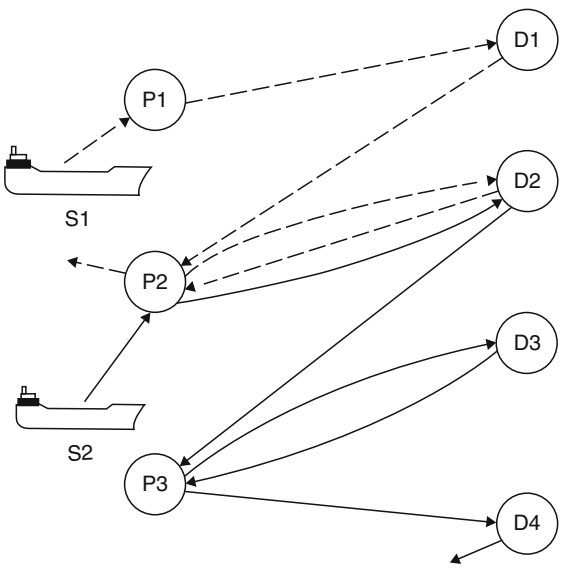


Fig. 3 Geographical routes for S1 (*dotted*) and S2 (*line*)

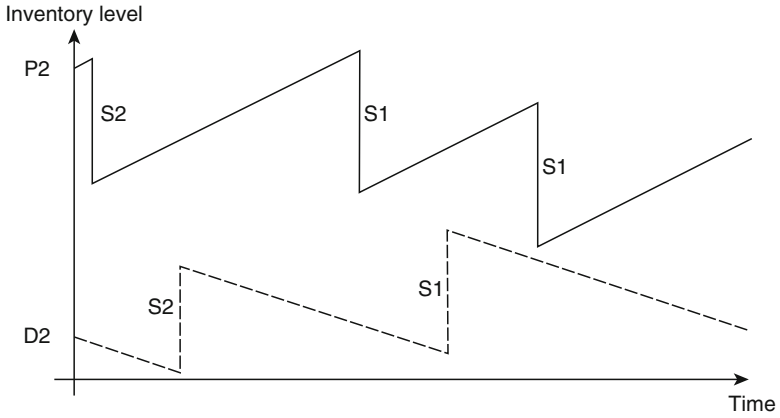


Fig. 4 Inventory levels for P2 (*line*) and D2 (*dotted*)

planning horizon. This means that the production at P2 probably is much larger than at P1 and/or the inventory capacity at P2 is very limited compared to the production.

Figure 4 shows the inventory levels for liquefaction plant P2 and regasification terminal D2 during the planning horizon. P2 is first visited by ship S2. Later in the planning horizon, ship S1 is visiting P2 twice. Note that the capacity of S1 is slightly larger than that of S2.

At the regasification terminals, the consumption level may vary throughout the planning horizon. However, in Fig. 4 constant consumption is assumed. There are numerous customers with different contracts at each regasification terminal. The sales contracts include fixed contracts where the agreed volume cannot be violated, contracts with lower and upper limits for the quantities to deliver, and short-term contracts which should be satisfied only if profitable. From this contract structure, lower and upper limits for gas demand with associated revenue can be specified for each time period. There are also given upper and lower limits on the inventory level.

The planning problem is to design routes and schedules including determining sales volumes that maximize the company revenue from the sales minus the operational costs. Moreover, neither the cargo capacities of the ships nor the inventory limits can be exceeded.

The problem is formulated as a path flow model, where the paths represent possible geographical routes for the LNG ships during the planning horizon. In addition, a path contains the schedule with information about the arrival times and the quantity unloaded at each regasification port call. Since we consider only full shiploads, the loading quantity is always equal to the capacity of the ship, while the unloading quantity at the regasification terminal is reduced due to boil-off.

In Fig. 3, we have two geographical routes with five calls for S1 and six calls for S2. In addition, there exists information about arrival times and loading and unloading quantities at each call on both these routes.

Table 1 Path information for S1

Geographical route	Arrival times at port (day)	(Un)loading quantity (m^3)
P1	T_1	V_v
D1	T_2	V_v–V_v $L(T_3–T_1)$
P2	T_3	V_v
D2	T_4	V_v–V_v $L(T_5–T_3)$
P2	T_5	V_v

For S1 in Fig. 3, a path could contain the information shown in Table 1. Here, T_k is the arrival time at port number k in the path and L is the boil-off parameter stating the share evaporating in each time period.

In addition to the notation defined in Sect. 3, we need the following. Let $\mathcal{R}_v$ be the set of paths for ship v indexed by r. Here, the path normally consists of several voyages in sequence, in contrast to the definition of $r \in \mathcal{R}_v$ in Sect. 4. The cost of sailing path r for ship v is given by the parameter C_{vr}. The cost parameters are composed of ship operations cost and port fees. Moreover, the parameter A_{ivrt} equals 1 if ship v visits port i in time period t on path r, and 0 otherwise. The unloading volume at the regasification terminals is given by the parameter Q_{ivrt}. The quantity sold at regasification terminal i at each time period t is regulated and must be within given lower and upper limits, $[D_{it}{}^{-}, D_{it}{}^{+}]$. The corresponding unit sales revenue is denoted by G_{it}. Finally, we need to define the path variable, λ_{vr} which is 1 if path r is sailed by ship v, and 0 otherwise, and the quantity of gas sold at regasification terminal i in time period t is d_{it}. Then, a path-flow formulation of the problem can be formulated as follows:

$$\max \sum_{i \in \mathcal{N}^D} \sum_{t \in \mathcal{T}} G_{it} d_{it} - \sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} C_{vr} \lambda_{vr}, \tag{5.1}$$

subject to

$$s_{it} - s_{i(t-1)} + \sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} V_v A_{ivrt} \lambda_{vr} = P_{it}, \quad \forall i \in \mathcal{N}^P, t \in \mathcal{T}, \tag{5.2}$$

$$s_{it} - s_{i(t-1)} - \sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} Q_{ivrt} \lambda_{vr} + d_{it} = 0, \forall i \in \mathcal{N}^D, t \in \mathcal{T}, \tag{5.3}$$

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} A_{ivrt} \lambda_{vr} \leq B_i, \quad \forall i \in \mathcal{N}, t \in \mathcal{T}, \tag{5.4}$$

$$S_i^- \leq s_{it} \leq S_i^+, \quad \forall i \in \mathcal{N}, t \in \mathcal{T}, \tag{5.5}$$

$$D_{it}^- \leq d_{it} \leq D_{it}^+, \quad \forall i \in \mathcal{N}^D, t \in \mathcal{T}, \tag{5.6}$$

$$\sum_{r \in \mathcal{R}_v} \lambda_{vr} = 1, \quad \forall v \in \mathcal{V}, \tag{5.7}$$

$$\lambda_{vr} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v. \tag{5.8}$$

The objective function (5.1) maximizes the revenue from the sales minus the transportation costs and the port fees. The inventory balances for the liquefaction plants are given in constraints (5.2). The inventory level at the end of time period t is equal to the inventory level at the end of the previous time period adjusted for the production and any loading in time period t. The loading quantity is determined by the capacity of the ship that is visiting the port in the particular time period. Similarly, constraints (5.3) give the inventory constraints for the regasification terminals. Constraints (5.4) make sure that the port capacity is not exceeded. Lower and upper bounds on the storage and sales variables are specified in constraints (5.5) and (5.6), respectively. Constraints (5.7) ensure that each ship sails exactly one path. Finally, the formulation involves binary requirements (5.8) on the ship path variables.

6 Solution Approaches

Both models can be solved by standard solvers for small instances only. The model in Sect. 4 is mainly used for annual planning and the numbers of potential voyages and δ_{vrt} variables are very high. Similarly, the paths and λ_{vr} variables in model (5.1)–(5.8) will be too many when solving the model by a commercial MIP solver for real-world instances. Therefore, such real problems have to be solved with adapted and often advanced solution approaches. These can be based on either exact solution methods or heuristics. We will here briefly present two solution approaches for the mathematical models presented in Sects. 4 and 5.

The models with a reduced number of voyage or path variables: Instead of generating all feasible voyage or path variables (columns), it is possible to generate a set of promising columns only. The challenge is then to find a set of promising columns that ensure a feasible and good solution. As an alternative, it is possible to start with a subset of all columns and add promising columns to the model during the solution process. Columns with negative reduced costs (for minimization problems) improve the LP-relaxed version (master problem) of the model and should be included in the model. When no potential columns can make improvements, the solution is optimal. To get integer solutions, a branch-and-bound algorithm can work with the columns generated for the first LP-problem only, or promising columns can be generated in each branch-and-bound node to find the optimal solution. The latter approach is called branch-and-price (Barnhart et al., 1998), and Grønhaug et al. (2010) use this approach for a real-world LNG inventory routing problem similar to the one described in Sect. 5. Computational results are given for instances with up to 5 ships, 6 ports, and 75 days planning horizon. Good feasible solutions were found for all instances, and for many of them the optimal solution was achieved.

The models with a reduced planning horizon: Here the whole planning horizon is divided into short time intervals, and each interval is then solved in sequence. Each

such interval spans a planning horizon that is so short that the problem can be solved by the models presented. First, the problem with the first and second time interval is solved. Then, the problem with the second and third time interval is solved, with the variables from the first time interval fixed to the values found when the first problem was solved. The solution process continues by including one new time interval and fixing the variables in the time intervals previously solved until the whole planning horizon is considered. This approach is often called a rolling horizon heuristic, and Moe et al. (2008) used this approach for a real large-scale problem similar to the one presented in Sect. 4. Here, an annual plan is developed for serving eight long-term contracts with monthly demands by use of more than 30 LNG ships.

7 Concluding Remarks

The LNG supply chain has become more complex in the last decade with the accelerating supply and demand for natural gas. The expanding opportunities in this business increase the need for and benefit from using models and tools to support the planning decisions. In this chapter, we have focused on the transportation planning and inventory management of the LNG supply chain. The routing of ships is closely related to the inventory management at the plants and terminals; hence, these planning aspects are considered simultaneously.

Two combined inventory management and transportation planning models are presented. The first model can typically be used within a gas producing company having one large production facility and many customers with delivery contracts. An example of such a producer is StatoilHydro with production from the gas field Snøhvit. The other model is well suited for a vertically integrated company with control of both the liquefaction plants and the regasification terminals in addition to the transportation. Several upstream suppliers and downstream customers are searching for such planning responsibility in the future even though they are not yet in that situation.

The first model can typically be used on a tactical planning level during negotiations about deliveries to different regasification terminals. An Annual Delivery Program (ADP) is created using the model, and the delivery dates at the regasification terminals are presented to the terminal owners. If some dates are inappropriate due to, for example, scheduled maintenance, this can be included in the model. When all delivery dates are approved, the ADP can be used in operation.

Since the second model includes sales, it is better suited for planning on an operational level. When the gas prices fluctuate, the model can be used to evaluate new sales opportunities and possibly reschedule ships for better economical utilization.

The models presented are simplified to ease readability. It is relatively easy to extend the models to include several qualities of LNG, several successive visits in the same type of port, spot and complicated contract considerations and price struc-

tures. However, some of these extensions would complicate the solution process and increase the running time of the models considerably.

In general, the models are very difficult to solve due to the combinatorial structure of the problems. The solution space is large with many feasible solutions to both models. Thus, these problems are very interesting for researchers within operations research, and there are many challenges within LNG supply chain optimization that are still unsolved.

Acknowledgment This work was partly supported by the OPTIMAR and DOMinant projects funded by the Research Council of Norway.

References

Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. P., & Vance, P. H. (1998). Branch-and-price: column generation for solving huge integer problems. *Operations Research, 46*(3), 335–348.

Christiansen, M., Fagerholt, K., Nygreen, B., & Ronen, D. (2007). Maritime transportation. In C. Barnhart, & G. Laporte (Eds.), *Handbooks in operations research and management science, vol. 14: transportation* (pp. 189–284). Amsterdam: North-Holland.

Christiansen, M., Fagerholt, K., & Ronen, D. (2004). Ship routing and scheduling: status and perspectives. *Transportation Science, 38*(1), 1–18.

EIA (2003). *The global liquefied natural gas market: status and outlook*. Washington, DC: Energy Information Administration, U.S. Department of Energy.

EnergyInfo. (2007). *The International Energy Outlook 2007. Resource document*. U.S. Energy Info. **Retrieved** April 24, 2008 from www.evworld.com/databases/shownews.cfm?pageid=news300302-01.

Grønhaug, R., & Christiansen, M. (2009). Supply chain optimization for the liquefied natural gas business. In L. Bertazzi, J. van Nunen, & M.G. Speranza (Eds.), *Innovation in distribution logistics, Lecture Notes in Economics and Mathematical systems* (pp. 195–218). Springer.

Grønhaug, R., Christiansen, M., Desaulniers, G., & Desrosiers, J. (2010). A Branch-and-price method for a liquefied natural gas inventory routing problem. Forthcoming in *Transportation Science*, DOI: 10.1287/trsc.1100.0317.

Moe, C. R., Rakke, J. G., & Stålhane, M. (2008). *Combined large scale LNG ship routing and inventory management*. Master's thesis, Norwegian University of Science and Technology.

NGR (2007). Natural Gas Market Review 2007. International Energy Agency.

Stremersch, G., Michalek, J., Hecq, S. (2008). *Decision support software tools for LNG supply chain management*. Gastech 2008.

Part IV
General Problems and Methods

Optimal Relinquishment According to the Norwegian Petroleum Law: A Combinatorial Optimization Approach

Horst W. Hamacher and Kurt Jörnsten

Foreword by Kurt Jørnsten

In the early 1990s, when I was on my way, on a Ruhrgas grant, to visit Horst Hamacher at the University of Kaiserslautern, my wife Åsa Hallefjord, who at that time was Head of Economics at Norske Shell AS, suggested that I and Horst should study the problem of relinquishment of oil licenses according to the Norwegian Petroleum Law. While in Kaiserslautern, Horst and I started to study the problem and tried to formulate it as a mathematical programming problem. However, very soon we ended up looking at a simpler but related problem, namely the k-cardinality tree problem. Soon, we also got Francesco Maffioli and Matteo Fischetti involved in the study of k-cardinality tree problems which resulted in the paper 'Weighted k-cardinality trees: complexity and polyhedral structure' published in *Networks*, 1994. The paper that Horst and I wrote on the original optimal relinquishment problem, however, remained a working paper[1] and was never published. This was partly due to the fact that the study of the k-cardinality tree problem got so interesting in itself and partly due to the fact that we did not manage to find a good mathematical programming formulation of the relinquishment problem. Since the study of the k-cardinality tree problem later led to more studies of combinatorial optimization problems with fixed cardinality constraints (as can be seen in the annotated bibliography on that topic published by Bruglieri et al. (2006)), it is of interest to make the paper on the original relinquishment problem available in printed format. Hence, we will here make the original optimal relinquishment more accessible to a wider audience, as the practical problem more or less generated a whole new area of study in combinatorial optimization: combinatorial optimization problems with

[1] See Jörnsten (1992).

H.W. Hamacher
Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany
e-mail: hamacher@mathematik.uni-kl.de

K. Jörnsten (✉)
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
e-mail: kurt.jornsten@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_25,
© Springer-Verlag Berlin Heidelberg 2010

fixed cardinality constraints. My wife Åsa Hallefjord who unfortunately died in a car accident in 1996 would have loved to see the problem of optimal relinquishment, according to the Norwegian Petroleum Law, that we introduced to have had such an impact on the development of the study of combinatorial optimization problems with fixed cardinality constraints. It is also interesting that, from a practical complexity issue, the Norwegian lawyers managed to select the value $N/2$ as the target value for relinquishment, thus creating a computationally very difficult problem to solve.

Abstract We present a combinatorial optimization model for the relinquishment of petroleum licenses in accordance with the rules given in the legislation concerning the Norwegian continental shelf. It is shown that a relaxation of the model can be interpreted as the problem of finding a minimal connected component of given cardinality in an undirected grid graph with node weights. We discuss the differences between the rules in 1973 and 1991, and how they lead to different mathematical models. A discussion of the complexity of the mathematical problem for optimal relinquishment is presented, and a presentation of related graph theoretical models is given. We conclude the paper with a small illustrative example, which shows the effects of the different restrictions. References to other studies in which large scale examples are treated are given.

1 Introduction

In the Act of 22 March 1985 No. 11 Pertaining to Petroleum Activities, it is stipulated in Section 1 that,

> "This Act applies to petroleum activities pertaining to exploration, exploration drilling, production, utilization, and pipeline transportation in Norwegian internal waters, in Norwegian territorial sea, and on the Continental Shelf. The Act also applies to installations for such activities and to shipment facilities for petroleum in the areas mentioned above. Installations do not comprise supply and standby vessels. The Act does not apply to movement of installations.
>
> The Act also applies to activities and installations as mentioned in first paragraph in areas outside the Continental Shelf to the extent such application follows from international law or from specific agreement with a foreign state.
>
> The King may issue regulations as to which provisions of this Act shall apply to pipelines with associated equipment in areas mentioned in first paragraph, when they are not owned by Norwegian licensees and neither start nor end within the said area.
>
> The Act does not apply to seabed subject to private property rights or to Svalbard's internal waters or territorial sea.
>
> As far as responsibility for pollution according to Chapter V and damage arisen as a result of pollution and waste according to Chapter VI special provision shall apply, cf. Section 38 and Section 44C second paragraph."

In this paper, we will be concerned with a specific question of great importance for a licensee on the Norwegian continental shelf, the question on how to obey the act's

demand for relinquishment. We will cite the section in the act that specifies the rules for relinquishment and other sections of importance for the relinquishment problem. We will then present a mathematical model for optimal relinquishment according to the act.

In Section 12, 'Area of a production licence', we are given the following information.

> "The areas mentioned in Section 1, first paragraph, shall be divided into blocks with size of 15 latitudal minutes and 20 longitudal minutes, unless adjacent land, boundaries with the continental shelves of other states, or other circumstances dictate otherwise. A production licence may comprise one or more blocks or part of blocks."

Section 13, 'The duration of a production licence', gives the following information.

> "The production licence shall be granted for up to 6 years. When justified by special reasons, the Ministry may extend the licence for periods of 1 year each up to a total of 4 years.
>
> A licensee who has fulfilled the work commitment imposed pursuant to Section 10 and the conditions which are otherwise applicable to the individual production licence, may demand that the licence shall continue to be valid for up to one half of the area originally covered by the licence, but not less than 100 sq. km, for a period to be stipulated in the individual licence, though not exceeding a further 30 years. If the licence applies to an area of less than 100 sq. km, an extension for the whole area may be demanded.
>
> The licence may be extended beyond the extension mentioned in second paragraph when justified by special reasons. The conditions for such special extension shall be stipulated by the Ministry. Application for extension must have been received 3 years prior to expiry of the licence.
>
> The King shall issue rules about how a demand for extension shall be presented and how areas to be relinquished shall be delineated."

In Regelverksamling for Petroleumsvirksomheten, 1991, 'Forskrifter til lov om petroleumsvirksomhet', the shape of the area to be relinquished is stipulated.

> "The areas to be relinquished by the licensee shall within each block be contiguous and limited by meridians and parallels of latitude expressed in whole minutes of a degree. No individual boundary lines shall be shorter than three latitudinal minutes and five longitudinal minutes. The area which the licensee may demand to retain shall be contiguous and shall amount to at least one-fifth of the initial area of the block unless otherwise agreed by the Norwegian Petroleum Directorate."

This is a strengthening of the requirements as compared with what is stipulated in the 'Legislation concerning the Norwegian continental shelf with unofficial English translation', 1973, based on Royal Decree of 8 December1972 relating to Exploration and Exploitation of Subsea Petroleum Reserves. In this document, the area to be relinquished is required to be continuous, whereas nothing is stated with respect to the continuity of the area to be retained. In Section 21 of the old law text the following is stated.

> "The areas which the licensee must relinquish pursuant to Section 20 shall form one continuous area bounded by lines of parallels of latitude and longitude expressed in whole degrees and whole minutes. No boundary line shall be less than 3 min of latitude and 5 min of longitude.

> The retained area of a block shall consist of at least 100 sq. km, unless otherwise approved by the Ministry.
>
> If discoveries of petroleum are made and such discoveries extend over an area exceeding half a block, the licensee may require that the prolongation shall apply to all deposits, even if this should result in retention of more than half the block by the licensee.
>
> In instances as described in the preceding paragraph, the Ministry shall fix the boundaries when the licensee has had an opportunity to present his view."

Going back to the act of 1985 and the rules from 1991, we are given the following information in Section 14, 'Relinquishment of production licence'.

> "The licensee may in the period mentioned in Section 13, first paragraph, upon 3 months notice, relinquish his rights pursuant to a production licence. Thereafter, relinquishment may take place at the end of each calendar year, provided notice of the relinquishment has been given at least 1 year in advance.
>
> If several licensees hold a production licence jointly, the Ministry may, under special circumstances, consent to the relinquishment by one or more of them of their shares in the licence.
>
> The relinquishment must apply to the whole area which is covered by the licence, unless the Ministry consents to that it shall only apply to part of the area."

Section 15, 'Area fee, production fee etc.', gives us details of costs for the licensee.

> "The licensee shall for the production licence, pay a fee per square kilometer (area fee) and a fee calculated on the basis of the quantity and value of petroleum produced at the shipment point of the production area (production fee).
>
> Nevertheless, production fee shall not be paid for petroleum produced from resources where the development and operation plan is approved after 1 January 1986.
>
> The Ministry may upon 6 months' notice decide that the production fee shall fully or partly be paid in form of petroleum produced. If required by the Ministry, the licensee shall ensure that this petroleum is transported, processed, stored, and made available at prices, priorities, and other conditions which are no less favourable than the terms applicable to the licensee's own petroleum from the relevant area. The Ministry may, upon 6 months notice, decide that the fee shall again be paid in cash.
>
> Petroleum which the state is entitled to receive as production fee, and the state's entitlement to transportation, processing, and storage of such petroleum may be transferred to others. Such transfer shall relieve the state of future obligations.
>
> The King may issue rules about the size of the fees and bonuses mentioned in first paragraph and about the calculation method, including provisions about stipulation of the value which shall form the basis for the calculation, about metering of the petroleum, and about information which the licensees shall provide about the production.
>
> The King may determine that the production fee fully or partly shall not be paid or that the duty to pay the production fee shall be postponed.
>
> Claims for fees, with addition of interest and costs, may be collected by distraint."

The size of the fees mentioned in Section 15, are specified in the rules, Regelverksamling for Petroleumsvirksomheten, 1991. In Section 12 of the rules, 'Area fees', we have the following.

"The licensee shall pay an area fee of NOK 3,000 per sq. km. for the period specified in Section 13, first paragraph, first sentence, of the Act. Upon extension pursuant to Section 13, first paragraph, second sentence, of the Act, a further NOK 3,500 per sq. km. per year shall be paid. The fee shall be paid in advance and will not be refunded even if the rights lapse before the period expires.

After expiry of the 6-year period, with possible extensions, an area fee shall be paid in advance for each calendar year according to the following: For the first year, NOK 5,500 per sq. km, shall be paid. Afterwards, the fee per sq. km, per year will increase to the following amounts:

Second year NOK 6,000

Third year NOK 7,500

Fourth year NOK 9,000

Fifth year NOK 15,000

Sixth year NOK 21,000

Seventh year NOK 27,000

Eighth year NOK 33,000

Ninth year NOK 39,000

Tenth year NOK 45,000

Thereafter the fee shall be NOK 90,000 per sq. km. per year for the remaining part of the duration of the licence. When calculating the fee, the area shall be rounded off to the nearest sq. km.

The Ministry may adjust the area fee at intervals of at least 5 years to bring it into line with changes in the value of the Norwegian kroner. The basis for the adjustment shall be the consumer price index of the Central Bureau of Statistics.

If the area fee is not paid when due, interest shall be payable on the amount due in accordance with the Act of 17 December 1976 No. 100 relating to Interest Overdue Payments."

The sections of the act pertaining to petroleum activities and the sections of the corresponding rules which we have cited give us the information needed to construct a mathematical model for the oil relinquishment problem facing licensees on the Norwegian continental shelf.

Before doing this, we will emphasize some changes in the rules by comparing the texts in the 1991 and the 1973 issue of the rules. The main difference is the stipulation of the retained area as given above. In the 1973 issue, Section 21, was

"The area which the licensee must relinquish pursuant to Section 20 shall form one continuous area bounded by lines of parallels of latitude and longitude expressed in whole degrees and whole minutes. No boundary line shall be less than 3 min of latitude and 5 min of longitude.

The retained area of the block shall consist of at least 100 sq. km, unless otherwise provided by the Ministry.

> If discoveries of petroleum are made and such discoveries extend over an area exceeding half a block, the licensee may require that the prolongation shall apply to all deposits, even if this should result in retention of more than half the block by the licensee.
>
> In instances as described in the preceding paragraph, the Ministry shall fix the boundaries when the licensee has had an opportunity to present his view."

As seen from the above, the new rules are a strengthening of the old rules. If these strengthened rules have had any effect on the way relinquishments have been made is beyond our knowledge. Looking at the relinquished areas in the North Sea, we can find examples where the retained areas are non-connected. However, as we will see, the strengthened rules have an impact on the mathematical model. To finalize this section we would also mention that a typical block in the North Sea has an area which is a little larger than 500 sq. km. Hence, the restrictions that the retained area shall be no less than one-fifth of the original area of the block or 100 sq. km, are approximately the same.

2 Constructing the Mathematical Model

The sections of the legal text cited above give us the constraints on our problem. Although the act contains information on situations in which one may overrule the strict rules given in the act and the accompanying set of rules, the licensee has to have good arguments to present to the Ministry in order to be able to keep an area which does not obey the rules in the legal text. Thus, a model which has the capabilities to analyze the effect of different relinquishment strategies can be a useful decision support.

What is the basis of knowledge on which the licensee shall take his decision regarding which areas to keep and which ones to be relinquished?

During the 6 years, the life time of the original licence, the licensee has conducted exploration activities. These activities consist of magnetic surveys, gravimetric surveys, seismic surveys, heat-flow measurements, radiothermic measurements, geochemical surveys, sampling of the seabed without drilling, and drilling to depth below seabed as stipulated in the individual case, but not deeper than 200 m. The licensee has also performed a work commitment program. This program may consist of the drilling of a certain number of wells down to specified depths or geological formations. The work program also includes core drillings and samplings as specified by the Norwegian Petroleum Directorate. Given that petroleum products are found, the information at hand also includes well testing and drilling of declination wells. This means that the licensee at the end of the 6-year period has more information on the hidden petroleum resources, their location, and production possibilities than in the start of the licence period.

Let us assume that the licensee uses this information to generate weights for each sub-block of the licence of size 1×1 min. Apart from the data obtained during the exploration program, these weights also include the fees as stated in the act and the

rules. Furthermore, let us assume that the original licence consisted of one block. This simplification is for ease of presentation only and can easily be relaxed. Note that the weights can be viewed as the net present value of the 1×1 min area of the block.

The oil relinquishment problem can now be formulated as a problem on a rectangular area consisting of 15×20 squares of size 1×1 min. Each square has an associated weight which gives the expected worth of the resources hidden under that area. We are now able to view the oil relinquishment problem as the problem of finding a connected region consisting of at least 150 squares and obeying the rules given in Section 21 of the legislation. For the licensee, optimal relinquishment means to find the connected region with minimum weight. This gives us the best area to relinquish.

The problem is illustrated in Fig. 1.

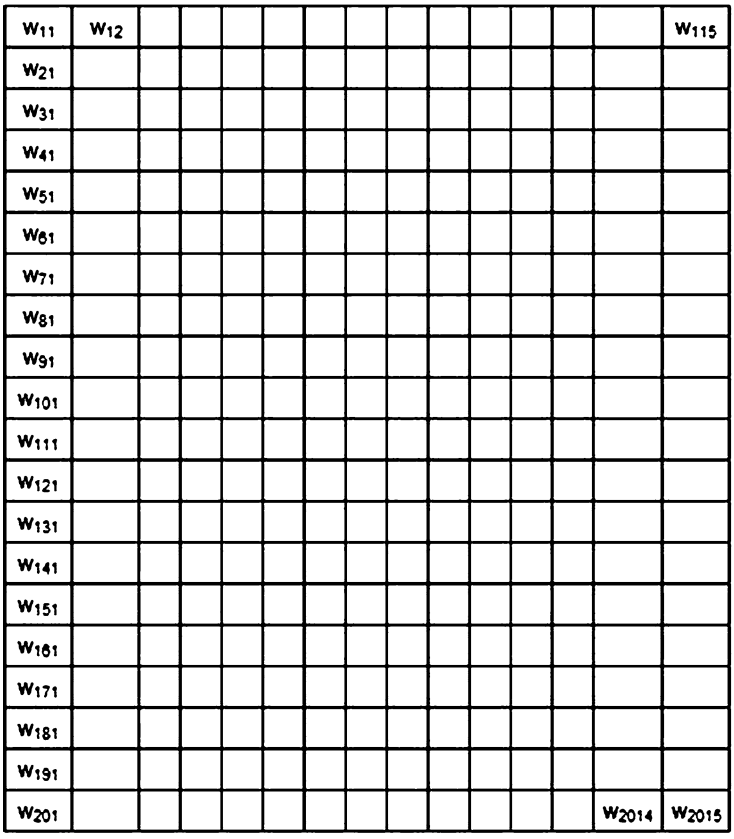


Fig. 1 In each square we are given a weight w_{ij}; all *empty squares* contain the corresponding w_{ij} value

3 Different Combinatorial Optimization Models

In the last paragraph, we have presented the relinquishment problem as a problem formulated on a rectangular area consisting of 300 unit squares. The problem is stated as selecting at least half of the squares such that the selected squares have minimal total weight, subject to constraints given in the legal text. In this form, it is difficult to see how these extra conditions can be formally stated. In what follows, we will give an alternative and more precise formulation of the relinquishment problem. This is done by formulating the relinquishment problem as a combinatorial optimization problem on an undirected graph.

Let us reformulate the relinquishment problem as a problem on an undirected grid graph of size 15×20 nodes. Each node has an associated weight w_{ij} and the undirected arcs connecting the neighboring nodes in the grid graph have arc weight zero. The oil relinquishment problem can now be stated as the problem of finding a connected node set with minimum weight and of cardinality at least ISO subject to the extra restrictions given by the minimal length on each subset defining the closure of the selected node set.

This is the formulation corresponding to the rules as given in the 1973 version. In the 1985/1991 version of the rules, the formulation has to be strengthened to a partition of the grid graph in two connected components, obeying the act's rules for minimal boundary lines.

We are now able to give an integer programming formulation of the relinquishment problem. As defined earlier let w_i denote the estimated value of the node i in the grid graph. Let Y_i be a binary variable associated with node i which takes the value one if node i is selected, zero otherwise. Let x_{ij} be a binary variable which takes the value one, if arc (i,j) is selected. Arc (i,j) connects nodes i and j. Note that we have used a simple index formulation, giving the 1×1 min areas, corresponding to nodes in the grid graph, consecutive numbers starting from the northwest corner and ending in the southeast corner. The numbering is assumed to be done row wise.

The oil relinquishment problem can now be stated as

$$\min_{x,y} \sum_{i \in N} w_i y_i$$

$$\text{subject to } \sum_{i \in N} y_i \geq \lfloor |N|/2 \rfloor$$

$$\sum_{(i,j) \in E} x_{ij} \geq \sum_{i \in N} y_i - 1$$

$$\sum_{(i,j) \in E(S)} x_{ij} \leq |S| - 1 \quad \forall S \subseteq N$$

$x_{ij} \leq y_i$ for all arcs (i,j) in E.

$$x_{ij} \leq y_j$$

The constraints stipulate that we must select at least $|N|/2$ nodes, or more specifically the integer part of $|N|/2$ nodes. The second constraint stipulates that we must select at least as many arcs in the grid graph G as the number of nodes selected minus one. This constraint together with the logical constraints that an arc can only be selected if the corresponding adjacent nodes have been selected guarantees that the selected sub-graph is connected. The simplest connected structure in a graph is a tree. In order to avoid that we get a result consisting of two or more connected sub-components, we add the third constraint. This constraint enforces a sub-tree of cardinality greater than or equal to $\lfloor |N|/2 \rfloor$. In the mathematical programming formulation given above we have not specified the requirements on the minimal length on the closure of the connected subset. However, these constraints can be stated as linear equations and inequalities in the variables x_{ij} and y_i.

The integer programming formulation given, states the problem according to the 1973 rules. Hence, the retained area that results from solving the model need not be connected. In order to have a connected retained region, the integer programming model above has to be expanded in order to guarantee that the retained region, corresponding to the variables $y_i = 0$ in the solution, also forms a connected component. We will not state the corresponding integer programming model here; however, by adding extra link variables z_{ij} which can only take the value one, if the corresponding y_i and y_j take the value zero, and adding constraint that stipulates that the graph structure in z has to form a tree, we have a valid model for the 1985/1991 rules. It should also be noted that the requirements given in the 1991 rules concerning connectivity in both the relinquished area and the retained area can be viewed as a graph partitioning problem in which the grid graph should be partitioned in two connected components, one of which specified by the requirement that it should contain at least $\lfloor |N|/2 \rfloor$ nodes and be of minimum weight. In a forthcoming paper, we will study the 1985/1991 problem and also discuss its relation to classical problems in graph partitioning. In the sequel, we will concentrate on the 1973 version of the problem unless otherwise specified.

4 Difficulty of the Relinquishment Problem

The mathematical programming problem given above has been studied by Hamacher et al. (1991) and by Fischetti et al. (1994). In both papers, the edge weight version of the problem is studied under the name weighted k-cardinality trees. In the 1991 paper, the following theorem is shown:

Theorem 1. *The minimum k-cardinality problem is strongly NP-hard.*

The proof is based on reduction to the Steiner tree problem.

Both Hamacher et al. (1991) and Fischetti et al. (1994) study the weighted k-cardinality problem on a general undirected graph. In Hijacker et al. a branch

and bound method is presented and some ways to obtain bounds are discussed. Fischetti et al. also contains a presentation of facets for the k-cardinality tree polytope. They also indicate how these facets can be used in a branch and cut algorithm.

Given the information that the k-cardinality tree problem is strongly NP-hard, we know that the problem is notoriously difficult to solve to optimality. Experiments with various heuristics and different branching schemes on some test problems have also shown that the problem is very difficult also in practical computational terms. Løkketangen and Jörnsten (1993, 1997) have used a tabu search strategy for the edge weighted k-cardinality tree problem. The results for relatively large graphs are satisfactory. At present, the same authors are working on a tabu search method for the 1973 version of the relinquishment problem. In this work, the authors also include the restrictions on the longitudal and latitudal boundaries given in the rules that accompany the act.

The k-cardinality tree in a grid graph has been studied by Woeginger (1992). Woeginger has shown that the edge weighted grid graph k-cardinality tree problem is NP-hard.

5 A Small Illustrative Example

Let us consider a small 4×4 example in which we are trying to find a connected region of size 8 of minimum weight. The weights for the 16 different sub-blocks are given in Fig. 2.

A solution in which we have selected the eight minimal values and relaxed the connectivity requirements is given in Fig. 3a. A solution in which the connectivity of the relinquished region is fulfilled is given in Fig. 3b. And finally in Fig. 3c, we give the solution for the problem in which also the retained area is connected. The example illustrates that the connectivity requirements have an effect even in a small scale example as the one given above. In order to illustrate the effect of the boundary line restrictions, an example of full size must be used.

9	6	8	0
9	9	9	8
7	5	9	8
1	7	7	9

Fig. 2 Weights for sub-blocks

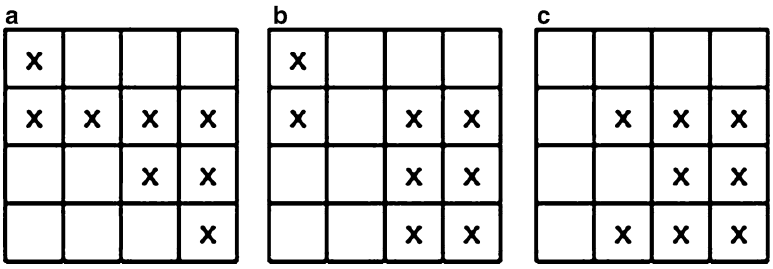


Fig. 3 (**a**) Solution with value 41, (**b**) Solution with value 43, and (**c**) Solution with value 45

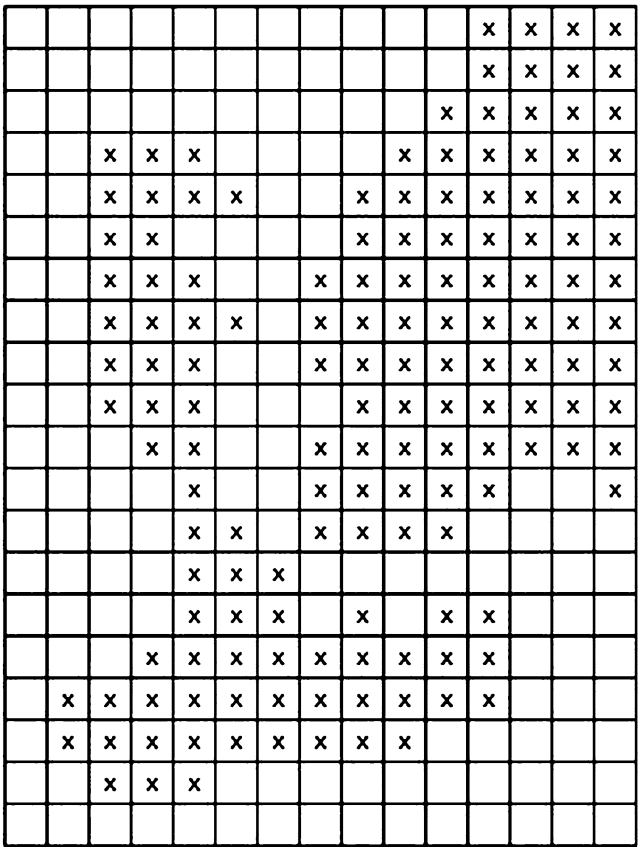


Fig. 4 Solution obeying connectivity rules; value of retained area 802

In all figures, the kept area is marked with x-s and hence the relinquished area is the blank area.

In Figs. 4–6, we give a full scale problem. We present the data for the potential and/or discovered petroleum reservoirs in the matrix $\mathrm{W} = (w_{ij})$. It should be

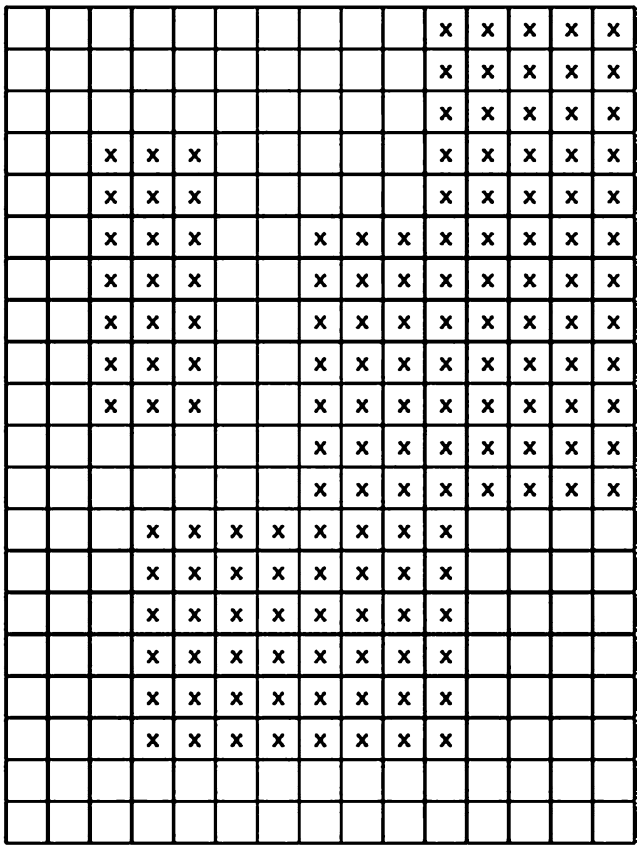


Fig. 5 Solution obeying 1973 rules. Value of retained area 752; retained area *shaded*

noted that the data used are generated so that the block shall contain three petroleum structures of different locations and extensions. Also it should be noted that the numbers used are far from being realistic. In a real life problem, many of the 1×1 min blocks would have negative values and the range of the worth for the specific 1×1 min blocks would surely be more diversified. Three solutions obeying only the connectivity requirement for the relinquished area, the 1973 and the1985/1991 rules accompanying the act are shown. The solution for the 1973 and the 1985/1991 rules are obtained using a heuristic as the solution methods for obtaining optimal solutions at present are incapable of dealing with problems of this size.

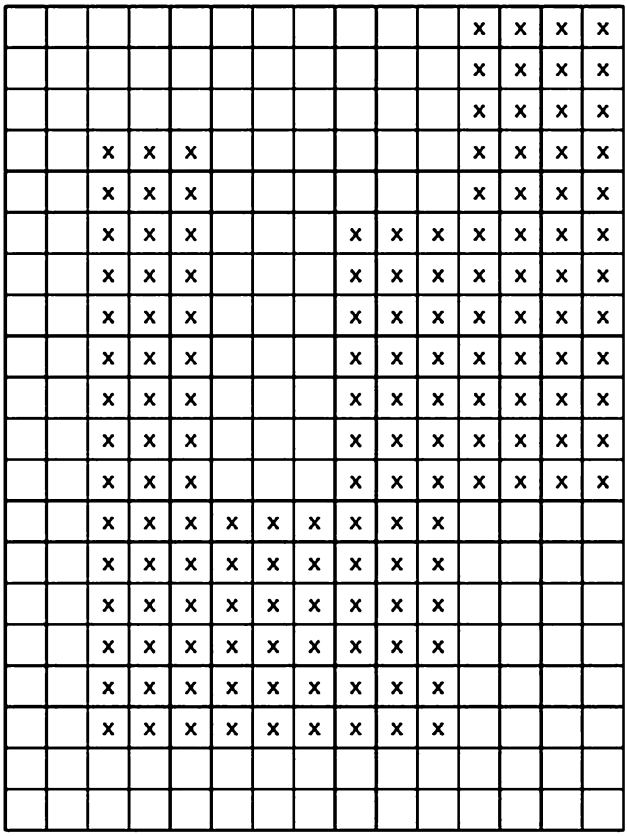


Fig. 6 Solution obeying 1985/1991 rules. Value of retained area 746; retained area *shaded*

The data for the problem of real life size are as follows:

$$w = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 4 & 5 & 7 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 5 & 6 & 8 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 3 & 4 & 6 & 7 & 9 \\
0 & 0 & 4 & 7 & 4 & 0 & 0 & 0 & 0 & 2 & 4 & 5 & 6 & 8 & 9 \\
0 & 1 & 5 & 8 & 4 & 2 & 0 & 0 & 2 & 3 & 4 & 5 & 7 & 8 & 10 \\
0 & 2 & 2 & 4 & 1 & 1 & 0 & 0 & 2 & 4 & 6 & 7 & 9 & 9 & 11 \\
0 & 1 & 4 & 3 & 2 & 1 & 0 & 2 & 3 & 5 & 7 & 7 & 9 & 9 & 10 \\
0 & 1 & 5 & 9 & 6 & 4 & 0 & 2 & 4 & 6 & 7 & 8 & 8 & 8 & 9 \\
0 & 1 & 7 & 8 & 7 & 2 & 0 & 3 & 4 & 6 & 7 & 8 & 8 & 8 & 9 \\
0 & 0 & 3 & 8 & 5 & 0 & 0 & 1 & 5 & 6 & 7 & 7 & 7 & 6 & 7 \\
0 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 6 & 6 & 7 & 6 & 6 & 6 & 6 \\
0 & 0 & 0 & 0 & 4 & 0 & 0 & 3 & 6 & 7 & 6 & 4 & 1 & 1 & 2 \\
0 & 0 & 0 & 0 & 3 & 6 & 1 & 2 & 4 & 3 & 3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 4 & 5 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 3 & 4 & 2 & 1 & 2 & 1 & 3 & 4 & 1 & 0 & 0 \\
0 & 0 & 1 & 2 & 8 & 6 & 6 & 3 & 4 & 3 & 2 & 5 & 1 & 0 & 0 \\
0 & 2 & 4 & 6 & 15 & 12 & 6 & 5 & 2 & 4 & 4 & 2 & 1 & 0 & 0 \\
0 & 2 & 6 & 7 & 13 & 10 & 5 & 6 & 4 & 2 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 3 & 4 & 5 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

In Fig. 4, we shade the retained region and as can be seen, the relinquished area is connected.

Note the solution also has a connected retained area if connectivity is defined as at least one common point.

6 Conclusions

We have presented the relinquishment problem facing licensees operating on the Norwegian continental shelf. We have presented combinatorial optimization formulations of different relinquishment problems and indicated their relations to different problems in graph theory. A discussion of the complexity of the relinquishment problem is given. References to related graph theoretical papers are given and also to a number of computational studies of related problems. The purpose of this paper has been to show how combinatorial optimization models can be used to formalize managerial problems arising from legal texts. It is interesting to see how Norwegian lawyers have been able to formulate a difficult and interesting graph theoretical problem. The work on the k-cardinality problem started in 1991 when the second author was visiting the University of Kaiserslautern on a Ruhrgas grant. All the studies on the edge weighted k-cardinality tree problem referred to in this paper have evolved from this point. According to our knowledge, these studies are the first studies of the weighted k-cardinality tree problem. We can thus state that the Norwegian lawyers that have written the act and the regulation rules have managed to open a new area of research in combinatorial optimization.

Acknowledgements We would like to thank Åsa Hallefjord, Head of Economics, Norske Shell AS for helping us to focus on the oil-license relinquishment problem.

References

Bruglieri, M., Ehrgott, M., Hamacher, H. W., & Maffioli, F. (2006). An annotated bibliography of combinatorial optimization problems with fixed cardinality constraints. *Discrete Applied Mathematics, 154*(9), 1344–1357

Fischetti, M., Hamacher, H. W., Jörnsten, K., & Maffioli, F. (1994). Weighted k-cardinality trees: Complexity and polyhedral structure. *Networks, 24*, 11–21

Hamacher, H. W., Jörnsten, K., & Maffioli, F. (1991). Weighted k-cardinality trees. Technical Report 91.023, Politecnico di Milano, Dipartimento di Elettronica, Milano, Italy

Jörnsten, K., & Løkketangen, A. (1997). Tabu search for weighted k-cardinality trees. *Asia-Pacific Journal of Operational Research, 14*(2), 9–26

Jörnsten, K., & Løkketangen, A. (1993). Tabu search for the optimal relinquishment problem. Research report, Møre og Romsda1 College, 1993

Jörnsten, K. (1992). *Optimal tilbakelevering av letelisenser*. Paper presented at FlBE (Fagkonferanse i Bedriftsøkonomiske Emner), January 1992, Bergen

Woeginger, G. J. (1992) Computing the optimum oil region in the plane. Research report, TU Graz, Institut fur Theoretische Informatik, Klosterwiesgasse, 31/11, A-8010 Graz, Austria

Acts and rules

Kongelig Resolusjon av 8.12.1972 om undersøkelse etter og utnyttelse av undersjøiske petroleumsforekomster. Royal Decree of 8.12.1972 relating to exploration and exploitation of subsea petroleum resources with unofficial English translation, 1973.

Lov om petroleumsvirksomhet, 1985. Stavanger: Oljedirektoratet. Act pertaining to petroleum activities (unofficial translation), 1985, Stavanger: Norwegian Petroleum Directorate.

Olje- og energidepartementet, (1991). Forskrifter til lov om petroleumsvirksomhet. Stavanger: Oljedirektoratet. Ministry of petroleum and energy, (1991). Regulations supplementing the act pertaining to petroleum activities (unofficial translation), Stavanger: Norwegian Petroleum Directorate.

An Overview of Models and Solution Methods for Pooling Problems

Dag Haugland

Abstract Network flow models, where the flow quality is given at the supply nodes and quality constraints are defined at the demand nodes, appear frequently when optimizing refinery operations, pipeline gas transportation, and other energy related operations. Tracing the quality of the flow from the supplier to the market implies that the quality must be updated at the nodes where different flow streams are mixed. This results in the well-known pooling problem, the non-linear nature of which makes it hard to solve. In this chapter, I give an overview of available methodology for solving the pooling problem. I distinguish between fast but possibly inexact methods aimed for medium and large scale instances, and more time-consuming exact methods that tackle small instances.

1 Introduction

Mathematical models for optimizing the product flow in energy production and transportation systems tend to increase dramatically in computational difficulty when the quality of the flow must be taken into account. Extending a single-commodity network flow model to a multi-commodity model may appear to be straightforward, but in cases in which the network nodes represent junction points where the commodities are blended, an otherwise linear planning model may have to be equipped with hard non-linear constraints. Particularly exposed to such characteristics are planning models designed for the oil refining industry. Here crude oils of various compositions constitute the flow, and storage or process units where crudes of unequal qualities are mixed represent the network nodes. Mixing of flow streams with possibly different compositions resulting in a new stream with an averaged composition is frequently referred to as pooling, and correspondingly, the planning models supporting this are known as *pooling problems*.

D. Haugland
Department of Informatics, University of Bergen, 5020 Bergen, Norway
e-mail: dag.haugland@ii.uib.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_26,
© Springer-Verlag Berlin Heidelberg 2010

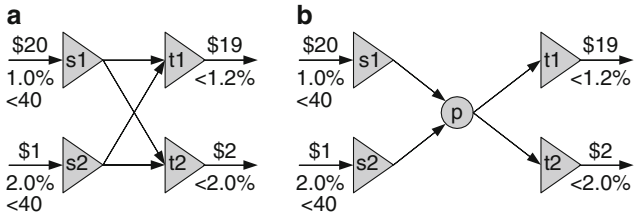


Fig. 1 A linear (**a**) and a non-convex (**b**) instance of the pooling problem

A simplistic version of the pooling problem is depicted in Fig. 1, which also demonstrates the transition from computationally easy, linear problems (Fig. 1a) to hard, non-linear (in fact non-convex) ones (Fig. 1b). The goal is to find the product flow from sources $s1$ and $s2$ to terminals $t1$ and $t2$ that maximizes the net revenues, given that the unit purchase costs at the sources are \$20 and \$1, respectively, while the unit sales prices at the terminals are \$19 and \$2, respectively. There are 40 units available at each source, and the relative contents of some chemical components are 1.0% and 2.0%, respectively. It is assumed that at each terminal, flow originating from different sources is mixed, and that the relative content of the component in question therefore becomes a weighted average of its content in the respective flow streams. The market at the two terminals will only accept to buy if the relative content is no larger than 1.2% and 2.0%, respectively.

The assumption in Fig. 1a is that each market can be supplied directly from each of the sources, whereas Fig. 1b depicts the unfortunate, but frequently occurring situation where some intermediate storage unit or some transportation unit must be shared between the flow streams. It is easily verified that in the first case, the solution is to supply $t1$ with a mix of 40 units from $s1$ and 10 units from $s2$, and send the remaining 30 units from $s2$ to $t2$, producing a net revenue of \$170. In the second case, the problem virtually is to decide what market the quality of the pool p should match. Satisfying the quality constraint at $t1$ implies that at least 80% of the pool feed must come from the purest source, $s1$. Purchasing 40 units from $s1$ and 10 units from $s2$, and sending it off to $t1$, hence yields a profit of \$140, which is the optimal solution. A pool quality of 2.0% means that only $t2$ can be supported, and the maximum profit is only \$40, which is obtained by feeding the pool exclusively from $s2$.

Note that the presence of the pool implies that the profit drops from \$170 to \$140. From a computational viewpoint, it is also interesting to observe that the decision problem in Fig. 1a can be modelled in terms of a *Linear Programme* (LP). This turns out not to be true for the second case. Since the exact solutions to large LP-instances can be found rapidly by the use of modern software packages, the type of problem illustrated in Fig. 1a represents no computational challenge, even if the numbers of sources, terminals and components subject to quality constraints grow very large. Due to its small size, the pooling problem depicted in Fig. 1b, could be solved by simple inspection. In general, however, one cannot expect the optimal solution to be found in reasonable computing time for all instance sizes that typically occur

in industrial applications. Consequently, the decision maker will have to rely on suboptimal solutions, like the alternative yielding only \$40 profit in the example above.

Since LP started to become a practical tool for optimized planning about half a century ago, the oil refining industry has been one of its major application areas. Pooling and blending are important processes, and quality constraints are often imposed on the content of contaminating components such as sulphur and CO_2. Consequently, the industry has for a long time recognized the challenge and the relevance of the pooling problem. Practical applications are demonstrated in e.g. Dewitt et al. (1989) and Baker and Lasdon (1985), and more recently in Amos et al. (1997).

The difficulty of reaching the optimal solution to the pooling problem, even in very small instances, was pointed out by Haverly (1978, 1979). Through numerical experiments, he proved that uncarefully designed methods may lead to solutions where only a small fraction of the acheivable profit is obtained. In realistic cases with a single pool, it is demonstrated that weak solution methods may imply that 75% of the profit is lost. To the oil refiner and the gas transporter, who so frequently are faced with planning problems involving pooled flow streams, efficient solution methods hence represent a significant economic potential.

The goal of this work is to give an overview of current solution methods for the pooling problem. I distinguish between *inexact* and *exact* methods. The former are supposed to be fast, but do not guarantee to locate the best solution. In contrast, exact methods have this capability, but may be prohibitively slow. The chapter is organized as follows. In the next section, I propose a general mathematical model covering a large class of pooling problems previously studied in the literature. This is used as a basis for a study of the categories of solution methods given in Sects. 3 and 4.

Throughout this text, it is assumed that the pool quality is a weighted average of the qualities of entering flow streams as explained in the example above. In some applications, however, this may be unrealistic. The octane number of a gasoline blend is for instance in general not given by such a formula, and the same applies to the combustive value of a composition of different gases. Taking into account the non-linear nature of the blending formulae that apply in such applications severely complicates the solution methods, and is defined beyond scope of the current work.

2 A General Pooling Model

Consider a directed graph $D = (N, A)$ with node and arc sets N and A, respectively, where $N = S \cup P \cup T$, and S, P, and T are the sets of sources, pools and terminals (sinks), respectively. Define L as a set of quality attributes or components. Furthermore, define the vectors $a \in \Re_+^S$ and $b \in \Re_+^T$ of supply and demand, the capacity and unit arc cost matrices $h, c \in \Re_+^A$, the source quality matrix $\lambda \in \Re^{S \times L}$, and the terminal quality bound matrix $d \in \Re^{T \times L}$.

With the exception of the two latter matrices, these are the data defining an instance of the *capacitated minimum cost flow problem*, on which the pooling model later is to be based. This problem amounts to allocate flow in the network in such a way that a_s units of flow enter the network at source $s \in S$, b_t units leave at terminal $t \in T$, flow is conserved at all nodes $p \in P$, and the total cost is minimized.

An additional constraint is that the content of component $l \in L$ at terminal $t \in T$ should not exceed d_t^l, given that it is λ_s^l in the supply to source $s \in S$, and that it is a weighted average of its content in entering streams at all nodes. To this end, define the decision variables $x \in \Re^A$ and $y \in \Re^{N \times L}$, representing flow and quality, respectively. Defining the sets $N^i = \{j \in N : (j,i) \in A\}$ and $N_i = \{j \in N : (i,j) \in A\}$ of upstream and downstream neighbours of any node $i \in N$, hence yields

$$y_i^l = \frac{a_i \lambda_i^l + \sum_{j \in N^i} x_{ji} y_j^l}{a_i + \sum_{j \in N^i} x_{ji}} \quad i \in N, \text{ where } a_i = \lambda_i^l = 0 \quad \forall i \in P \cup T, l \in L.$$

The general definition of the pooling problem becomes:

$$\min \sum_{(i,j) \in A} c_{ij} x_{ij} \tag{1}$$

$$\sum_{j \in N^i} x_{ji} - \sum_{j \in N_i} x_{ij} = \begin{cases} -a_i, & i \in S \\ 0, & i \in P \\ b_i, & i \in T \end{cases} \tag{2}$$

$$a_i y_i^l + y_i^l \sum_{j \in N^i} x_{ji} - \sum_{j \in N^i} x_{ji} y_j^l = a_i \lambda_i^l, \quad i \in N, l \in L \tag{3}$$

$$y_t^l \le d_t^l \quad t \in T, l \in L \tag{4}$$

$$0 \le x_{ij} \le h_{ij} \quad (i,j) \in A. \tag{5}$$

Problem (1)–(5) is a straightforward extension of the formulation by Foulds et al. (1992), which by introduction of multiple qualities (the set L) was made slightly more general by Adhya et al. (1999). Both cited works assume a network in three layers of nodes (sources, pools and sinks), where any arc goes from a source to a pool, from a pool to a sink, or from a source directly to a sink. What is new here is the absence of such restrictions on network topology, and thus e.g. arcs between two sources or two pools are allowed.

All quality constraints (4) are given as upper bounds, but it is easily seen that if there is a lower bound on the content of some component, this can be incorporated within (1)–(5). The model is also general in the sense that it covers the situation where either supply or demand is given as either a lower or an upper bound. In the cases of lower supply bounds and upper demand bounds, I now go on to show how the network must be modified in order to transform the problem to the form (1)–(5). For the minimum cost flow problem, similar transformations introducing a super

source and a super sink τ exist. The transformations all extend T by τ, but in order to take source quality into account, it may be necessary to introduce more than one new source.

2.1 Lower Supply Bound

Consider some $s \in S$ where the supply constraint $\sum_{j \in N^s} x_{js} - \sum_{j \in N_s} x_{sj} \geq a_s$ applies. Define the new arcs (s', s) and (s', τ), where s' is a new supply node with fixed supply $\sum_{j \in N_s} h_{sj} - a_s$ and quality $\lambda^l_{s'} = \lambda^l_s \ \forall l \in L$. By fixing the supply at s to a_s, it is ensured that exactly a_s units with quality λ^l_s enter the network at s, and via s' more flow with the same quality may enter if this is optimal.

2.2 Upper Demand Bound

Consider some $t \in T$ where the demand constraint $\sum_{j \in N^t} x_{jt} - \sum_{j \in N_t} x_{tj} \leq b_t$ applies. This requires the introduction of one new source s' and one new terminal t'. The demand at t' is fixed to b_t, whereas the demand at t is redefined to be zero. As shown in Fig. 2, arcs (t, t'), (s', t'), and (s', τ) are introduced. The quality constraints at t are kept unchanged, and the quality bounds at t' are defined as $d^l_{t'} = \lambda^l_{s'} = d^l_t$.

The units entering the network at s' can now partly flow along (s', t') to reach t', and partly be discarded by flowing along (s', τ). By constraint (2) applied to s' and t', $x_{s'\tau} = x_{tt'}$, and this quantity represents the actual delivery at t, whereas $x_{s't'}$ equals the corresponding slack in the demand constraint at t. Although t gets zero external flow after the modification, it is kept as a demand node in order to preserve

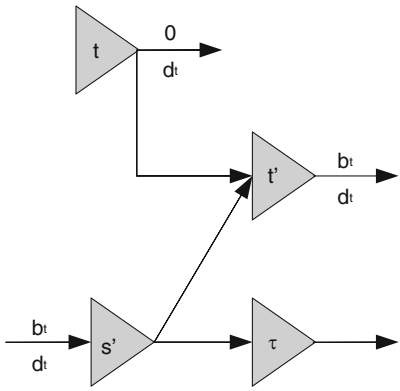


Fig. 2 Network updates for upper demand bounds

the effect of its quality bounds on downstream neighbour nodes. Likewise, it is not possible to introduce an arc directly from s' to t. This would degrade the flow at t if t receives flow with better quality than d_t^l from its upstream neighbours. Hence the new terminal t' is required.

2.3 Generality of the Model

It is left to the reader to find the network modifications transforming upper supply bounds and lower demand bounds to a formulation on the form (1)–(5). For the super-sink τ, define the demand such that total demand in all sinks equals total supply in all sources, and put $d_\tau^l = \max\left\{\lambda_i^l : i \in S\right\}\ \forall l \in L$. Without exclusion of important special cases, this model thus can be adopted as a definition of the pooling problem.

The computationally problematic part of the problem is the non-linear nature of constraints (3). By fixing either the flow or the quality variables, (3) become linear and are consequently referred to as *bilinear* constraints. Methods suggested for the pooling problem merely vary in how they deal with bilinearities, and discussions on this are given in the next two sections.

3 Inexact Solution Methods

Without comparison, the most frequently used method to tackle pooling problems in commercial applications, is by some use of *successive linear programming* (SLP). The main idea behind this method is to replace each bilinear term xy (for convenience, subscripts and superscripts are omitted) in constraints (3) by the tangent plane of the function $f(x, y) = xy$ at a guessed solution (x', y'). Hence the linear term $y'x + x'y - x'y'$ appears as a replacement for xy, which renders (1)–(5) a LP. This is exploited in an iterative manner until the guessed solution turns out to be the optimal solution of the LP:

1. Guess the optimal value of all flow and quality variables.
2. Replace bilinear terms and solve the resulting LP.
3. If for some variable the LP-solution differs from the guessed value, then let the LP-solution be the new guess, and go to Step 2.
4. Otherwise, output the LP-solution and stop.

As pointed out by e.g. Haverly (1978), SLP is not even guaranteed to converge if implemented in such a simple manner as indicated above. This is accounted for by bounding the length of the move away from the guessed solution, and by shrinking the bound when signs of divergence are detected (Baker and Lasdon 1985).

SLP dates back to the early work of Griffith and Stewart (1961). Although few theoretical results on SLP are available, the method gained popularity through

good numerical performance. Low implementation cost due to extensive reuse of LP-code is probably another reason why it became the preferred approach for tackling nonlinear elements in otherwise linear models. However, the frequently cited experiments of Haverly (1978, 1979) demonstrate that the method is sensitive to the initial guess, and that there is a considerable risk of being trapped in suboptimalities. This triggered the interest in more carefully designed SLP-techniques, and later also methods that guarantee the optimal solution.

With the purpose of faster convergence and better likelihood of locating the optimal solution, variants and improvements of SLP have been suggested in e.g. Palacios-Gomez et al. (1982), Zhang et al. (1985), and Sarker and Gunn (1997).

To reduce the effect of the initial guess, the idea of repeating SLP for a set of different initial solutions suggests itself. Audet et al. (2004) suggest a heuristic method for doing this in a randomized way. When the first SLP has been solved, the solution is slightly perturbed at random, and this perturbed solution becomes the next initial guess. If the new SLP-solution is better than the current, it replaces the current solution, and the above is repeated. Otherwise, the current solution is kept, and the amount by which the solution can be perturbed is increased. This procedure is repeated a fixed number of iterations.

The SLP technique used in the heuristic of Audet et al. (2004) is a simplification of the general one. Rather than using $y'x + x'y - x'y'$ as replacement for xy, the linear approximations $x'y$ and $y'x$ are used in an alternating manner. This corresponds to a technique often referred to as recursion (Haverly 1978; Main 1993), which in early applications has been preferred to SLP.

4 Exact Solution Methods

Although it is commonly recognized that in many practical applications, the problem size permits only inexact methods, most of the research in the last two decades has been focused on exact ones. I distinguish between approaches based on *flow* formulations such as (1)–(5), and approaches based on *proportion* formulations.

4.1 Flow-Based Methods

Most exact methods for the pooling problem are based upon *relaxations* as opposed to approximations of the bilinear constraints. A relaxation is a sub-problem with the important property that its feasible region contains all feasible solutions to the original problem. Computationally efficient relaxations are also characterized by a convex and smallest possible feasible region. To this end, the concepts of convex and concave *envelopes* of a function f on a convex set Ω become relevant. These are functions of the same variables as f, defined as $vex_{\Omega} f(x) = \sup_{g \in V(f,\Omega)} g(x)$

and $cav_\Omega f(x) = \inf_{h \in C(f,\Omega)} h(x)$, respectively, where $V(f,\Omega)$ is the set of convex functions g that for all $y \in \Omega$ satisfy $g(y) \leq f(y)$, and $C(f,\Omega)$ is the set of concave functions h that for all $y \in \Omega$ satisfy $h(y) \geq f(y)$. Hence the best convex relaxation of the constraint $f(x) = 0$, $x \in \Omega$, is $vex_\Omega f(x) \leq 0 \leq cav_\Omega f(x)$.

McCormick (1976) showed that if $f(x,y) = xy$ and $\Omega = [\underline{x}, \overline{x}] \times \left[\underline{y}, \overline{y}\right]$, where $\underline{x}$, $\overline{x}$, $\underline{y}$ and $\overline{y}$ are fixed lower and upper bounds on x and y, respectively, then

$$vex_\Omega f(x,y) = \max\left\{\underline{y}x + \underline{x}y - \underline{x}\,\underline{y}, \overline{y}x + \overline{x}y - \overline{x}\,\overline{y}\right\}, \tag{6}$$

$$cav_\Omega f(x,y) = \min\left\{\underline{y}x + \overline{x}y - \overline{x}\,\underline{y}, \overline{y}x + \underline{x}y - \underline{x}\,\overline{y}\right\}. \tag{7}$$

Furthermore, $vex_\Omega f(x,y) = xy = cav_\Omega f(x,y)$ if $x = \underline{x}$, $x = \overline{x}$, $y = \underline{y}$ or $y = \overline{y}$. Al-Khayyal and Falk (1983) utilized (6)–(7) in an algorithm for minimizing a bilinear function subject to linear constraints, and proved that the algorithm converged to the optimal solution to this nonconvex problem. This is also the basis of the exact pooling problem algorithm developed by Foulds et al. (1992). Assume finite lower and upper bounds on all variables are available (preprocessing can easily yield such bounds). Each bilinear term xy is then replaced by a new variable ξ along with the constraints $\xi \geq \underline{y}x + \underline{x}y - \underline{x}\,\underline{y}$, $\xi \geq \overline{y}x + \overline{x}y - \overline{x}\,\overline{y}$, $\xi \leq \underline{y}x + \overline{x}y - \overline{x}\,\underline{y}$, and $\xi \leq \overline{y}x + \underline{x}y - \underline{x}\,\overline{y}$.

The resulting relaxation is an LP. If for all bilinear terms, either of the two variables has optimal LP-value equal to one of its bounds, the optimal solution to the pooling problem is found. Otherwise, the algorithm picks one bilinear term xy for which this is not the case, and defines four new sub-problems where (x,y) is bounded within $[\underline{x}, x'] \times \left[\underline{y}, y'\right]$, $[x', \overline{x}] \times \left[\underline{y}, y'\right]$, $[\underline{x}, x'] \times [y', \overline{y}]$ and $[x', \overline{x}] \times [y', \overline{y}]$, respectively. Here (x', y') denotes the value of (x,y) in the LP-solution. Resembling the *branch-and-bound* algorithm for integer programmes, the method of Foulds et al. (1992) applies the procedure above recursively to each new sub-problem, and stops the recursion when the LP-solution is feasible in the pooling problem or it has higher cost than some known feasible solution.

Audet et al. (2000) have later improved this algorithm by applying e.g. a *branch-and-cut* technique, and in Audet et al. (2004), this method was successfully applied to instances where e.g. $|N| = 15$ and $|L| = 4$.

4.2 A Proportion Model

For a given network structure, Ben-Tal et al. (1994) suggested an alternative formulation for the pooling problem that does not require explicit quality variables.

Assume that D has no directed paths with more than two arcs. That is, each arc in D either goes from a source to a pool, from a pool to a terminal, or directly from a source to a terminal. For each $s \in S$ and $p \in P$, define the variable

$q_{sp} = \frac{x_{sp}}{\sum_{i \in N^p} x_{ip}} = \frac{x_{sp}}{\sum_{j \in N_p} x_{pj}}$. That is, q_{sp} denotes the proportion of the flow at p that enters via arc (s, p).

Replacing x_{sp} by $q_{sp} \sum_{t \in N_p} x_{pt}$ for all $p \in P$, $(s, p) \in A$, y_p^l by $\sum_{s \in N^p} \lambda_s^l q_{sp}$ for all $p \in P$, $l \in L$, and y_t^l by $\frac{\sum_{p \in P \cap N^t} \left(x_{pt} \sum_{s \in N^p} \lambda_s^l q_{sp} \right) + \sum_{s \in S \cap N^t} \lambda_s^l x_{st}}{b_t}$ for all $t \in T$, $l \in L$, problem (1)–(5) can be written:

$$\min \sum_{s \in S} \sum_{p \in P \cap N_s} c_{sp} \sum_{t \in N_p} q_{sp} x_{pt} + \sum_{t \in T} \sum_{j \in N^t} c_{jt} x_{jt} \tag{8}$$

$$\sum_{p \in P \cap N_s} q_{sp} \sum_{t \in N_p} x_{pt} + \sum_{t \in T \cap N_s} x_{st} = a_s, \; s \in S \tag{9}$$

$$\sum_{j \in N^t} x_{jt} = b_t, \quad t \in T \tag{10}$$

$$\sum_{p \in P \cap N^t} \left(x_{pt} \sum_{s \in N^p} \lambda_s^l q_{sp} \right) + \sum_{s \in S \cap N^t} \lambda_s^l x_{st} \le b_t d_t^l, \quad t \in T, \quad l \in L \tag{11}$$

$$q_{sp} \sum_{t \in N_p} x_{pt} \le h_{sp}, \quad p \in P, \quad s \in N^p, \tag{12}$$

$$\sum_{s \in N^p} q_{sp} = 1, \quad p \in P, \tag{13}$$

$$q_{sp} \ge 0, \quad p \in P, \quad s \in N^p, \tag{14}$$

$$0 \le x_{jt} \le h_{jt}, \quad t \in T, \quad j \in N^t. \tag{15}$$

The proportion model (8)–(15) corresponds to the so-called q-formulation in Ben-Tal et al. (1994). Multiplying (13) by x_{pt} yields the valid inequality $x_{pt} = \sum_{s \in N^p} q_{sp} x_{pt}$. Sahinidis and Tawarmalani (2005) suggested adding this equality to the q-formulation, producing the pq-formulation, and demonstrated that the relaxation thus becomes stronger (its feasible region decreases).

Audet et al. (2004) provided an experimental comparison between the flow model and the q-formulation by applying their branch-and-cut algorithm to both, and found that the latter could solve larger instances than could the flow model. Instances that could be solved by applying any of the two models were solved faster by using the q-formulation.

By applying their generic global optimization code, BARON (Sahinidis 1996), to the *pq*-formulation, Sahinidis and Tawarmalani (2005) managed to solve instances where $|N| = 15$ and $|L| = 6$ with very small search trees and CPU-time less than 2 s on a 1.7 GHz Pentium IV processor. To the best of our knowledge, this represents the best performance on exact solution of pooling problems to date.

5 Conclusions

When taking the composition of the flow in a network into consideration, the otherwise linear network model are likely to be turned into hard pooling problems. In, for instance, gas transportation and the processing industry, notably in oil refining, the pooling problem finds numerous applications. Due to its bilinear constraints, it represents a challenge to the optimizer, which traditionally has been met by inexact successive linear programming methods.

When designing exact methods, finding a formulation with a strong linear relaxation is crucial. The first such models that were suggested consist of flow and quality variables. Replacing the quality variables by variables representing flow proportions have later been shown to perform better when fed into generic branch-and-cut algorithms.

Extending the applicability of proportion models indicate a possible direction for further research. As pointed out by Audet et al. (2004), the *q*- and *pq*-formulations are not directly applicable if D contains paths of more than one pool. The authors also indicate a hybrid model, with both quality and proportion variables. Following Main (1993), the restriction in network structure is not desirable, since paths of pools make it even more difficult for SLP-techniques to find the optimum. Such networks do have practical relevance. Consider a multi-period inventory model, where the new supply is mixed with what currently is stored in the inventory, and quality of the inventory input varies over time. In the corresponding network, there is a path of pools for each inventory, and the lengths of these paths equal the number of periods. In forthcoming research on the pooling problem, it should be studied how generalized versions of the *pq*-formulation can be applied in order to tackle networks with paths of multiple pools.

References

Adhya, N., Tawarmalani, M., & Sahinidis, N. V. (1999). A Lagrangian approach to the pooling problem. *Industrial and Engineering Chemistry Research*, *38*(5), 1956–1972

Al-Khayyal, F. A., & Falk, J. E. (1983). Jointly constrained biconvex programming. *Mathematics of Operations Research*, *8*(2), 273–286

Amos, F., Rönnqvist, M., & Gill, G. (1997). Modelling the pooling problem at the New Zealand Refining Company. *Journal of the Operational Research Society*, *48*(8), 767–778

Audet, C., Brimberg, J., Hansen, P., Le Digabel, S., & Mladenović, N. (2004). Pooling problem: Alternate formulations and solution methods. *Management Science*, *50*(6), 761–776

Audet, C., Hansen, P., Jaumard, B., & Savard, G. (2000). A branch-and-cut algorithm for non-convex quadratically constrained quadratic programming. *Mathematical Programming*, *87*(1), 131–152

Baker, T. E., & Lasdon, L. S. (1985). Successive linear programming at Exxon. *Management Science*, *31*(3), 264–274

Ben-Tal, A., Eiger, G., & Gershovitz, V. (1994). Global minimization by reducing the duality gap. *Mathematical programming*, *63*(2), 193–212

DeWitt, C. W., Lasdon, L. S., Waren, A. D., Brenner, D. A., & Melhem, S. A. (1989). OMEGA: An improved gasoline blending system for Texaco. *Interfaces*, *19*(1), 85–101

Foulds, L. R., Haugland, D., & Jörnsten, K. (1992). A bilinear approach to the pooling problem. *Optimization*, *24*, 165–180

Griffith, R. E., & Stewart, R. A. (1961). A nonlinear programming technique for the optimization of continuous processing systems. *Management Science*, *7*, 379–392

Haverly, C. A. (1978). Studies of the behavior of recursion for the pooling problem. *ACM SIGMAP Bulletin*, *25*, 19–28

Haverly, C. A. (1979). Behavior of recursion model – more studies. *ACM SIGMAP Bulletin*, *26*, 22–28

Main, R. A. (1993). Large recursion models: Practical aspects of recursion techniques. In T. A. Ciriani & R. C. Leachman (Eds.), *Optimization in Industry* (pp. 241–249). New York, USA: John Wiley & Sons Ltd

McCormick, G. P. (1976). Computability of global solutions to factorable non-convex programs: Part I – convex underestimating problems. *Mathematical Programming*, *10*(1), 147–175

Palacios-Gomez, F., Lasdon, L. S., & Engquist, M. (1982). Nonlinear optimization by successive linear programming. *Management Science*, *28*(10), 1106–1120

Sahinidis, N. V. (1996). BARON: A general purpose global optimization software package. *Journal of Global Optimization*, *8*(2), 201–205

Sahinidis, N. V., & Tawarmalani, M. (2005). Accelerating branch-and-bound through a modeling language construct for relaxation specific constraints. *Journal of Global Optimization*, *32*(2), 259–280

Sarker, R. A., & Gunn, E. A. (1997). A simple SLP algorithm for solving a class of nonlinear programs. *European Journal of Operational Research*, *101*(1), 140–154

Zhang, J., Kim, N. -H., & Lasdon, L. S. (1985). An improved successive linear programming algorithm. *Management Science*, *31*(10), 1312–1331

Cooperation Under Ambiguity*

Sjur Didrik Flåm

Abstract Exchange of contingent claims is construed here as a cooperative game with transferable utility. Solutions are sought in the core. The novelty is that agents, being uncertainty averse, may use distorted, subjective probabilities. Choquet integrals therefore replace expected utility. When convoluted payoff is concave at the aggregate endowment, there is a price-generated, explicit core solution.

1 Introduction

By a *contingent claim* is meant a mapping $x : S \to \mathcal{X}$ from a suitable scenario or *state set* S into a standard, finite-dimensional *commodity space* $\mathcal{X}$. In state $s \in S$, the owner of x can claim commodity bundle $x(s) \in \mathcal{X}$.[1] Suppose contingent claims can be added and scaled; then they form a linear space $\mathbb{X} \subseteq \mathcal{X}^S$.

Considered below is voluntary exchange of such claims. Taking part in the exchange are economic agents mentioned on a fixed, finite list A. Those agents are seen as profit-oriented producers – or as consumers who enjoy quasi-linear utility. Accordingly, agent $a \in A$ worships maximization of his pecuniary payoff function $\Pi_a : \mathbb{X} \to \mathbb{R} \cup \{-\infty\}$. By assumption, he already owns "endowment" $e_a \in \mathbb{X}$.[2] So, he might contend with taking home payoff $\Pi_a(e_a)$. Most likely, however, he can do better.

*Dedicated to Kurt Jørnsten on his 60th birthday.

[1]Typically, $\mathcal{X} = \mathbb{R}^G$, where G denotes a finite, fixed set of goods. Finance and insurance deals with instances that feature merely one good, namely money. More general examples include state-contingent ownership of different natural resources – or of permits to emit diverse pollutants.

It is tacitly assumed that the state, unknown ex ante, becomes common knowledge ex post. Otherwise information is *asymmetric*; see Flåm and Koutsougeras (2010).

[2]A producer's endowment is identified with his state-dependent bundle of output commitments or claims to production factors.

S.D. Flåm
Department of Economics, University of Bergen, 5020 Bergen, Norway
e-mail: sjur.flaam@econ.uib.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_27,
© Springer-Verlag Berlin Heidelberg 2010

To wit, because claims and payoffs are perfectly transferable, a coalition $C \subseteq A$ could aim at overall payoff

$$\Pi_C(e_C) := \sup \left\{ \sum_{a\in C} \Pi_a(x_a) \;\middle|\; \sum_{a\in C} x_a = \sum_{a\in C} e_a =: e_C \right\}. \tag{1}$$

$Sup\,\{\cdot\}$ is shorthand for *supremum* – that is, the least upper bound – of the bracketed set. Clearly, $\sup\{\cdot\} = \max\{\cdot\}$ iff that bound belongs to the set $\{\cdot\}$.

Since $\Pi_C(e_C) \geq \sum_{a\in C} \Pi_a(e_a)$, coordination (1) offers advantages – especially for large coalitions. So, a chief issue is whether the total payoff to the *grand coalition* $C = A$ can be realized and shared in ways compatible with the incentives of various parties. Reflecting on this issue, we declare a cash payment scheme $\kappa = (\kappa_a) \in \mathbb{R}^A$ in the *core*[3] iff it entails

$$\left.\begin{array}{ll} \textit{Pareto efficiency:} & \sum_{a\in A} \kappa_a = \Pi_A(e_A), \text{ and} \\ \textit{no blocking:} & \sum_{a\in C} \kappa_a \geq \Pi_C(e_C) \text{ for each coalition } C \subset A. \end{array}\right\} \tag{2}$$

The subsequent analysis revolves around core solutions. It starts from the observation that contingent claims and endowments are often traded. In particular, the setting suits reinsurance and security markets where parties pursue monetary objectives.[4]

In such markets, however, participants – or their consultants – hardly state and solve problems like (1). Rather, solutions and equilibria, if any, typically emerge in a decentralized manner, via exchange governed by market-clearing prices. Accordingly, with a view towards the first fundamental welfare theorem, we seek core solutions that comply with competitive pricing in exchange economies.

Maintaining this market-orientation, the paper offers two novelties. First, it goes beyond *risk* by tolerating *uncertainty*. Second, it employs more general decision criteria than expected utility.[5] The extension derives from merely assuming that every agent is uncertainty averse, hence has a taste for hedging. Formally, the representation of such attitudes amounts to replace probability measures with

[3] The *cooperative game* at hand has *player set* A and so-called *characteristic function* $C \mapsto \Pi(e_C)$.

[4] In case of risk, typically described by common or objective probabilities, several studies have already dealt with Pareto efficient allocations. But clearly, besides being efficient, participation had better be voluntary as well. Thus the *core* naturally occupies center stage; see Baton and Lemaire (1981); Borch (1960a,b).

[5] The paradigm of expected utility still holds dominant sway in economic theory. It has, however, been provoked by empirical paradoxes and challenged by concerns with the foundations of Bayesian decision making; see Dempster (1968), Machina (1987), and Schmeidler (1989). To mitigate matters, numerous generalized criteria have come up during the last decades; see Fishburn (1988) for review.

convex capacities and expectations with non-additive integrals of Choquet type, as explained in Sect. 3.[6]

The paper addresses diverse readers, concerned with finance, insurance or cooperative game theory. It invites consideration – or brief tasting – of some selected subjects from the said fields. Therefore, the paper should not be regarded as a fully spelled out, didactical text. It rather seeks to offer some publicity for optimization theory, especially Lagrangian duality, as well as for Choquet integrals. Transferable utility serves well in this regard. It obviates fixed-point arguments, opens for convex analysis or tractable computation, and it underscores the prominent role of comonotonicity.

Section 2 recalls how shadow prices generate core solutions. Section 3 reviews Choquet integration and prepares for Sect. 4 to outline that uncertainty-averse preferences typically admit integral representations. Sections 5 and 6 bring the preceding objects together and study specific properties of core solutions.

2 Shadow Prices and Core Solutions

This section sets the stage and is self-contained. It briefly reviews how and why shadow prices suffice to generate core solutions. For generality in argument and simplicity in exposition, specification of the objective $\Pi_a(\cdot)$ is deferred until Sect. 4. For the same reason, at this stage, it facilitates discussion to regard endowments and contingent claims simply as vectors in a real linear space $\mathbb{X}$.

As said, our aim is to find computable core solutions, generated by prices. To that end, write $\mathbf{x} = (x_a) \in \mathbb{X}^A$ for a profile $a \in A \mapsto x_a \in \mathbb{X}$. Further, let $x^* : \mathbb{X} \to \mathbb{R}$ be any linear functional, and associate to problem (1) its standard *Lagrangian*

$$L_C(\mathbf{x}, x^*) := \sum_{a \in C} \Pi_a(x_a) + x^* \left(\sum_{a \in C} e_a - \sum_{a \in C} x_a \right).$$

It makes for easier notation to write simply $x^* x$ instead of $x^*(x)$.

Definition. (Shadow price) *Any linear* $\lambda : \mathbb{X} \to \mathbb{R}$ *such that* $\sup_{\mathbf{x}} L_A(\mathbf{x}, \lambda) \leq \Pi_A(e_A)$ *will be named a Lagrange multiplier or* ***shadow price****.*

To appreciate such multipliers, imagine a perfectly competitive market for vectors in $\mathbb{X}$ that features linear price x^*. Clearly, having free and voluntary access to a market of that sort cannot harm anybody. Therefore, optimal transactions, if any, yield

[6] This extension retains the elegance and tractability of the von Neumann–Morgenstern–Savage model. In addition, it can accommodate bid-ask spreads, belief functions, distorted probabilities, evidence weights, non-linear pricing, sunspots, and omitted states; see Gilboa (1987), Gilboa and Schmeidler (1994), Karni and Schmeidler (1991), Schmeidler (1986), Schmeidler (1989), and Shafer (1976). It has already been applied to Pareto optimal sharing; see Chateauneuf et al. (2000) and insurance premia; see Wang et al. (1997), but not to cooperative contracts as here.

$$\sup_{\mathbf{x}} L_C(\mathbf{x}, x^*) = \sup_{\mathbf{x}} \left\{ \sum_{a\in C} \Pi_a(x_a) + x^* \left(\sum_{a\in C} e_a - \sum_{a\in C} x_a \right) \right\}$$
$$\geq \Pi_C(e_C) \text{ for each } C \subseteq A.$$

In particular, $\sup_{\mathbf{x}} L_A(\mathbf{x}, x^*) \geq \Pi_A(e_A)$. Thus, $x^* = \lambda$ is a shadow price iff

$$\sup_{\mathbf{x}} L_A(\mathbf{x}, x^*) = \Pi_A(e_A),$$

meaning that the said market equilibrates. Now note that for any linear x^*,

$$\sup_{\mathbf{x}} \left\{ \sum_{a\in C} \Pi_a(x_a) + x^* \left(\sum_{a\in C} e_a - \sum_{a\in C} x_a \right) \right\} = \sum_{a\in C} \left\{ \Pi_a^{(*)}(x^*) + x^* e_a \right\}$$

where the convex function

$$x^* \mapsto \Pi_a^{(*)}(x^*) := \sup \{ \Pi_a(x_a) - x^* x_a \mid x_a \in \mathbb{X} \}$$

is called the *conjugate* of Π_a. The economic meaning of conjugation is straightforward: If producer a obtains revenue $\Pi_a(x_a)$ from factor input x_a, and takes the factor price x^* as given, $\Pi_a^{(*)}(x^*)$ denotes his maximal profit. In view of (2), the preceding arguments prove the following.

Theorem 2.1. (Shadow prices support core solutions, Evstigneev and Flåm (2001)) *Let λ be any shadow price. Then the payment scheme that offers agent $a \in A$ the cash amount*

$$\kappa_a(\lambda) := \Pi_a^{(*)}(\lambda) + \lambda e_a, \tag{3}$$

constitutes a core solution.

Observe that (3) pays each agent in two capacities: first, $\Pi_a^{(*)}(\lambda)$ as production profit and second, λe_a for his endowment. Also note that core solution (3) is generated by a price λ that fully decentralizes individual choice.

At this juncture, the reader may rightly wonder whether shadow prices exist. After all, many markets tend to operate inefficiently. Particularly problematic are economies of scale, featuring non-concave objectives. So, precisely where and how does concavity enter the scene?

To elucidate that issue it is expedient to regard shadow prices as *marginal payoffs* – that is, as "*gradients*", brought to the fore by differential calculus. For the statement, given a function $f : \mathbb{X} \to \mathbb{R} \cup \{-\infty\}$, a linear mapping $x^* : \mathbb{X} \to \mathbb{R}$ is called a *supergradient* of f at x, and we write $x^* \in \partial f(x)$, iff $f(x)$ is finite, and

$$f(\hat{x}) \leq f(x) + x^*(\hat{x} - x) \text{ for all } \hat{x} \in \mathbb{X}.$$

Two features of this concept must be underscored. First, from a local perspective, if f is classically differentiable at x, then $\{x^*\} = \partial f(x)$ is a singleton,

containing only the customary gradient $x^* = f'(x)$. Second, in a more important and far-reaching optic, each supergradient $x^* \in \partial f(x)$ yields affine and global support of f from above – with coincidence (tangency) at x.

In terms of such generalized gradients, what comes next is a crucial characterization of shadow prices. For brevity, declare an allocation $(x_a) \in \mathbb{X}^A$ *feasible* iff $\sum_{a\in A} x_a = e_A$, and *optimal* if $\sum_{a\in A} \Pi_a(x_a) = \Pi_A(e_A)$ as well.

Theorem 2.2. (Shadow prices as supergradients)

- *λ is a shadow price iff $\lambda \in \partial\Pi_A(e_A)$. Thus, given the individual criteria, a shadow price depends only on the aggregate endowment e_A.*
- *For any $\lambda \in \partial\Pi_A(e_A)$, alongside an optimal allocation (x_a), we have $\lambda \in \partial\Pi_a(x_a)$ for all a. Conversely, if some λ belongs to each $\partial\Pi_a(x_a)$, and $\sum_{a\in A} x_a = e_A$, then λ is a shadow price, and allocation (x_a) is optimal.*

Proof. The assertions are well known when all functions Π_a are concave; see Laurent (1972). Here, however, concavity is not presumed. So, some extra work is needed. For simplicity, define the "death" penalty $\delta(\cdot)$ on $\mathbb{X}$ by $\delta(x) = +\infty$ when $x \neq 0$ and $\delta(0) = 0$. Note that this function has Fenchel conjugate $\delta^*(x^*) := \sup_x \{x^* x - \delta(x)\} \equiv 0$. Now $\lambda \in \partial\Pi_A(e_A)$

$$
\begin{aligned}
&\Leftrightarrow \sum_{a\in A} \Pi_a(x_a) - \delta\left(\sum_{a\in A} x_a - x\right) \leq \Pi_A(x) \\
&\quad \leq \Pi_A(e_A) + \lambda(x - e_A)\ \forall x \in \mathbb{X}, \forall (x_a) \in \mathbb{X}^A \\
&\Leftrightarrow \sum_{a\in A} \{\Pi_a(x_a) + \lambda(e_a - x_a)\} + \lambda\left(\sum_{a\in A} x_a - x\right) - \delta\left(\sum_{a\in A} x_a - x\right) \\
&\quad \leq \Pi_A(e_A)\ \forall x, \forall (x_a) \\
&\Leftrightarrow \sum_{a\in A} \{\Pi_a(x_a) + \lambda(e_a - x_a)\} + \delta^*(\lambda) \leq \Pi_A(e_A)\ \ \forall (x_a) \in \mathbb{X}^A \ \ (*) \\
&\Leftrightarrow \sup_{\mathbf{x}} L_a(\mathbf{x}, \lambda) \leq \Pi_A(e_A).
\end{aligned}
$$

This proves the first bullet. For the second, let $(\tilde{x}_a)$ be any optimal allocation. In the above string of equivalences, line (*) says

$$
\begin{aligned}
\lambda \in \partial\Pi_A(e_A) &\Leftrightarrow \sum_{a\in A} \Pi_a(x_a) \leq \sum_{a\in A} \{\Pi_a(\tilde{x}_a) + \lambda(x_a - \tilde{x}_a)\}\ \ \forall (x_a) \in \mathbb{X}^A \\
&\Leftrightarrow \Pi_a(x_a) \leq \Pi_a(\tilde{x}_a) + \lambda(x_a - \tilde{x}_a)\ \forall x_a \in \mathbb{X}, \forall a \Leftrightarrow \lambda \in \partial\Pi_a(\tilde{x}_a)\ \forall a. \ \square
\end{aligned}
$$

Many aspects of Theorem 2.2 merit mention. Clearly, the exchange market, whether fictitious or real, cannot balance unless $\lambda \in \partial\Pi_A(e_A)$. Further, Pareto efficiency cannot prevail unless marginal payoffs are equal, meaning $\lambda \in \cap_{a\in A} \partial\Pi_a(x_a)$ for any optimal allocation (x_a). In short, these optimality conditions are necessary. But it is most crucial that, thanks to global support, they become sufficient as well.

Finally, note that concavity is not really necessary; neither Π_a nor Π_A need be concave. What imports is merely to have Π_A supported from above at e_A by some affine function.

Taken together, these observations beg questions as to the availability of shadow prices. On that account Theorem 2.2 immediately yields.

Proposition 2.1. (Existence of shadow prices) *Denote by* $\hat{\Pi}_A : \mathbb{X} \to \mathbb{R} \cup \{-\infty\}$ *the pointwise smallest concave function* $\geq \Pi_A$. *Suppose*

$$\hat{\Pi}_A(\cdot) \text{ has a supergradient at } e_A \text{ with } \hat{\Pi}_A(e_A) = \Pi_A(e_A). \tag{4}$$

Then there exists at least one shadow price. Conversely, if a shadow price exists, then (4) holds.

Also noteworthy is the validity of *the law of demand*:

Proposition 2.2. ("Price curves slope downwards") *For any two shadow prices* $\lambda \in \partial\Pi_A(e_A),\ \hat{\lambda} \in \partial\Pi_A(\hat{e}_A)$ *it holds that* $(\lambda - \hat{\lambda})(e_A - \hat{e}_A) \leq 0$.

Proof. $\lambda \in \partial\Pi_A(e_A)$ implies $\Pi_A(\hat{e}_A) \leq \Pi_A(e_A) + \lambda(\hat{e}_A - e_A)$. Quite similarly, $\hat{\lambda} \in \partial\Pi_A(\hat{e}_A)$ implies $\Pi_A(e_A) \leq \Pi_A(\hat{e}_A) + \hat{\lambda}(e_A - \hat{e}_A)$. Add the two inequalities to conclude. □

Since price curves are paramount, it should comfort followers of Adam Smith – and students of microeconomics – that adding many and small agents to A makes the convoluted payoff $\Pi_A(\cdot)$ increasingly concave; see Evstigneev and Flåm (2001), Flåm et al. (2005). Broadly, provided agents be numerous and negligible, a shadow price is likely to exist.

Anyway, granted existence, it facilitates analysis that all objectives admit tractable representations. In this regard, much convenience already derives from construction (1) by way of being separable across agents. Additional tractability obtains when individual objectives $\Pi_a(\cdot)$ are separable over states as well. That issue is studied in the subsequent two sections.

3 Choquet Integration

This section takes time out to consider the Choquet integral and merely *one agent*. Moreover, to fit finance and insurance, only *one commodity,* referred to as *money,* will come into consideration.[7]

From here on, the state set S is equipped with a sigma-algebra Σ of observable events. $\mathbb{M}$ *denotes the linear space of all bounded,* Σ*-measurable mappings*

[7] More precisely: until Sect. 5, posit $\mathcal{X} = \mathbb{R}$.

$m : S \to \mathbb{R}$.[8] It is understood that $m(s)$ denotes money received in state s. For any event $\mathcal{E} \in \Sigma$, the *indicator function* $1(\mathcal{E}, \cdot) = 1_{\mathcal{E}} : S \to \mathbb{R}$ equals 1 on $\mathcal{E}$ and 0 elsewhere.

Suppose the agent at hand evaluates monetary claims $m \in \mathbb{M}$ by averaging payments somehow – that is, by using some sort of *generalized integral* $I(m)$, may be not quite standard. Outlined next is a chief generalization, namely the *Choquet integral*.

For the argument, let $m \in \mathbb{M}$ first be *simple*, meaning that its real values r_k, $k = 1, \ldots, K$ are finitely many and different. Posit $\mathcal{E}_k := \{s \in S \mid m(s) = r_k\}$, and $S_k := \mathcal{E}_1 \cup \ldots \cup \mathcal{E}_k$ to get

$$m = \sum_{k=1}^{K} r_k 1(\mathcal{E}_k) = \sum_{k=1}^{K} r_k \left[1(S_k) - 1(S_{k-1})\right] = \sum_{k=1}^{K} (r_k - r_{k+1}) 1(S_k), \tag{5}$$

where S_0 is empty, and $r_{K+1} := 0$. Now, in case the integral I equals the customary expectation $\mathbb{E}_P$, taken with respect to a probability measure P on Σ, the linearity of $\mathbb{E}_P$ yields:

$$I(m) = \mathbb{E}_P m = \sum_{k=1}^{K} r_k P(\mathcal{E}_k) = \sum_{k=1}^{K} (r_k - r_{k+1}) P(S_k).$$

Further, when $m \geq 0$, suppose, with no loss of generality, that $r_1 > \cdots > r_K \geq r_{K+1} = 0$ to have $S_k = \{s \mid m(s) > r_{k+1}\}$. Hence, the preceding equation reads:

$$\mathbb{E}_P m = \sum_{k=1}^{K} P\{m > r_{k+1}\} (r_k - r_{k+1}). \tag{6}$$

For any *non-negative* $m \in \mathbb{M}$, upon approximating it by a sequence of simple, non-negative functions, the limit of (6) gives the well-known formula

$$\mathbb{E}_P m = \int_0^{+\infty} P\{m > r\}\, dr. \tag{7}$$

For general $m \in \mathbb{M}$ it holds, because of its boundedness, that $m + \rho \geq 0$ provided the constant ρ be sufficiently large. So, formula (7) implies $\mathbb{E}_P m = \mathbb{E}_P(m + \rho) - \rho =$

$$\int_0^{+\infty} P\{m + \rho > r\}\, dr - \rho = \int_{-\rho}^{0} (P\{m > r\} - 1) dr + \int_0^{+\infty} P\{m > r\}\, dr.$$

Because the last right-hand side is unaffected by letting $\rho \to +\infty$, formula (7), valid only for $m \geq 0$, generalizes for arbitrary $m \in \mathbb{M}$ to

$$\mathbb{E}_P m = \int m dP = \int_{-\infty}^{0} (P\{m > r\} - 1) dr + \int_0^{+\infty} P\{m > r\}\, dr, \tag{8}$$

[8] For simplicity or computation, one may envisage S as finite - and let Σ comprise all its subsets. Then *every* $m : S \to \mathbb{R}$ is Σ-measurable.

or equivalently, to the more familiar and symmetric formula

$$\mathbb{E}_P m = -\int_{-\infty}^{0} P\{m \le r\}\, dr + \int_{0}^{+\infty} P\{m > r\}\, dr,$$

valid whenever m has a mean. After so much review of the standard expectation $\mathbb{E}_P$, we return to our agent. May be he cannot quite settle on *one* probability measure P. In fact, he could face some ambiguity as to which measure might prevail. To wit, suppose his beliefs be embodied in a set function $c : \Sigma \to \mathbb{R}$, called *capacity*, that reflects various degrees of confidence, likelihood or weight. The capacity ought, however, resemble a probability somehow. So, as a minimum, tacitly assume c *normalized*: $c(\emptyset) = 0$, $c(S) = 1$, and *monotone*: $\mathcal{E} \subset \mathcal{E}' \Rightarrow c(\mathcal{E}) \le c(\mathcal{E}')$.

Then, upon replacing P with c on the right-hand side of (8), what obtains is called the *Choquet integral with respect to capacity* c, defined by

$$\int m dc := \int_{-\infty}^{0} (c\{m > r\} - 1) dr + \int_{0}^{+\infty} c\{m > r\}\, dr. \tag{9}$$

As introduced here, the Choquet integral is well defined,[9] and it underscores the importance of how function values are ranked:

Example 3.0. (*Choquet integral of simple functions*) Suppose $m \in \mathbb{M}$ has finitely many real values $r_1 < \cdots < r_k$. With $\{m > r\} := \{s \in S \mid m(s) > r\}$, formula (9) gives

$$\int m dc = r_1 + (r_2 - r_1) c\{m > r_1\} + \cdots + (r_K - r_{K-1}) c\{m > r_{K-1}\}. \ \diamond \tag{10}$$

Most important, unlike $\mathbb{E}_P$, the Choquet integral need not be additive. Then, what sort of weaker additivity might reasonably hold? For an answer, recall that $I(m) = \int m dc$, in the face of uncertainty, represents the agent's ex-ante valuation of monetary claim m. Up front, naturally suppose he *prefers hedging* in the sense that

$$I(m + m') \ge I(m) + I(m') \ \text{ for all } m, m' \in \mathbb{M}. \tag{11}$$

To see what inequality (11), called *superadditivity*, means for the underlying capacity, posit $m = 1_{\mathcal{E}}$ and $m' = 1_{\mathcal{E}'}$ for events $\mathcal{E}, \mathcal{E}' \in \Sigma$. Since $m + m'$ then takes values in $\{0, 1, 2\}$, whatever be the properties of c, it follows from (10) that

$$\int (1_{\mathcal{E}} + 1_{\mathcal{E}'}) dc = c(\mathcal{E} \cup \mathcal{E}') + c(\mathcal{E} \cap \mathcal{E}').$$

[9] The integrands on the right hand side of (9) are monotone decreasing. The integral there is the classical one of Riemann–Stiltjes.

In particular, for $\mathcal{E} = \mathcal{E}'$, we get $\int 1_{\mathcal{E}} dc = c(\mathcal{E})$, this telling that the Choquet integral regenerates its underlying capacity. Further, under superadditivity (11),

$$c(\mathcal{E} \cup \mathcal{E}') + c(\mathcal{E} \cap \mathcal{E}') \geq c(\mathcal{E}) + c(\mathcal{E}'). \tag{12}$$

A capacity c that satisfies (12) is declared *supermodular* or *convex*.[10]

Example 3.1. (*Convex distortions*) Convex capacities obtain as convex distortions $c = d \circ P$ of a probability measure P. That is, $c(\cdot) = d(P(\cdot))$, where the "distortion function" $d : [0,1] \to [0,1]$ is convex, and $d(0) = 0$, $d(1) = 1$; see Denneberg (1994). $\diamond$

For economists, the word *convexity* indicates that the margin increases. That word is appropriate here because (12) is equivalent to

$$c(\mathcal{E}' \cup \Delta) - c(\mathcal{E}') \geq c(\mathcal{E} \cup \Delta) - c(\mathcal{E}) \text{ whenever } \mathcal{E} \subseteq \mathcal{E}' \text{ and } \mathcal{E}' \cap \Delta = \varnothing.$$

The preceding arguments prove half of the

Theorem 3.1. (On superadditivity, Denneberg (1994)) *The agent prefers hedging iff the underlying capacity is convex.*[11]

Some additional properties of the Choquet integral are worth recording. For one, return to the simple situation where m has values $r_1 > \cdots > r_K$, and $\mathcal{E}_k = \{s \mid m(s) = r_k\}$. Let π denote any permutation of $1, \ldots, K$ and posit $S_{\pi,k} := \mathcal{E}_{\pi 1} \cup \ldots \cup \mathcal{E}_{\pi k}$, $S_{\pi,0} := \varnothing$. Consider the (linear programming) problem

$$\text{minimize } \sum_{k=1}^{K} r_{\pi k} P_\pi(\mathcal{E}_{\pi k}) \text{ over permutations } \pi$$
$$\text{when } P_\pi(\mathcal{E}_{\pi k}) := c(S_{\pi k}) - c(S_{\pi(k-1)}).$$

Now, c being convex, one may show that $P_\pi \geq c$ on all events defined in terms of $\mathcal{E}_1, \ldots, \mathcal{E}_K$; see Denneberg (1994). Of particular importance here is the measure P_{id}, where $id(k) = k$ for each k. Indeed, to minimize the above sum one should match the largest value r_1 with the smallest margin $c(S_1) - c(S_0) = P_{id}(\mathcal{E}_1)$, the second largest value r_2 with the second smallest margin $c(S_2) - c(S_1) = P_{id}(\mathcal{E}_2)$, and so on. Put differently: the larger the value, the smaller the corresponding probability. These observations go some way to justify.

Theorem 3.2. (On convex capacity and pessimism) *The capacity c is convex iff*

$$\int m dc = \min \{\mathbb{E}_P m \mid \textit{probability measure } P \geq c\}. \tag{13}$$

(13) mirrors some *pessimism* in so far as always using a worst probability $P \geq c$.[12]

[10] For set functions these notions originated in cooperative game theory; consult Osborne and Rubinstein (1994).

[11] That is, the Choquet integral is superadditive iff the capacity is supermodular.

[12] When c is convex, there does indeed exist a probability measure $P \geq c$; consult Shapley (1971).

Examples 3.2. (*Finite state set*) Let c here be convex, S finite, and Σ contain all subsets. By (13), calculating $\int mdc$ amounts to solve the linear program

$$\text{minimize} \sum_{s\in S} m(s)P(s) \text{ s.t. } \sum_{s\in S} P(s) = 1 \text{ and } P(S') \geq c(S') \text{ for all } S' \subset S.$$

(*On fuzzy beliefs over two outcomes*) In particular, when S is a two-point set $\{s, \hat{s}\}$, convexity means $c(s) + c(\hat{s}) < 1$, see Chateauneuf et al. (2000), Dow et al. (1992), Tallon (1998). Then, with $c(s) \leq c(\hat{s})$, any probability measure $P \geq c$ corresponds, in one-to-one manner, to a point probability $p(s) \in [c(s), 1 - c(\hat{s})]$ on state s. Thus, (13) amounts to

$$\int mdc = \min\{c(s)m(s) + (1 - c(s))m(\hat{s}), (1 - c(\hat{s}))m(s) + c(\hat{s})m(\hat{s})\}.$$

(*On complete ambiguity*) In case $c(S') = 0$ for each proper subsets of $S' \subset S$, *every* probability measure P is $\geq c$ whence formula (13) gives $\int mdc = \min_{s\in S} m(s)$. Note that *risk aversion*, alias *diminishing marginal utility*, did not enter here. It was rather total uncertainty about P, and attending preference for hedging, that brought the worst outcome into focus. $\diamond$

It follows from (9) that the Choquet integral is *monotone*:

$$m \leq \bar{m} \Rightarrow \int mdc \leq \int \bar{m}dc, \tag{14}$$

but, as said, it is not generally additive. There is, however, an important exception:

Proposition 3.3. (Comonotone additivity, Denneberg (1994)) *It holds*

$$\int (m + \bar{m})dc = \int mdc + \int \bar{m}dc \;\; \textit{whenever} \;\; m, \bar{m} \;\; \textit{are} \;\; \textbf{\textit{comonotone}}, \tag{15}$$

meaning

$$\big[m(s) - m(s')\big] \cdot \big[\bar{m}(s) - \bar{m}(s')\big] \geq 0 \;\; \textit{for all pairs of states} \;\; s \neq s'. \tag{16}$$

Example 3.3. (*Comonotonicity, Choquet integral, and ordering of values*) Let m be simple with distinct values $r_1, \ldots, r_K$, and posit $S_k := \{s \mid m(s) \geq r_k\}$. Then, provided $r_1 > \cdots > r_K$, the indicators $1(S_k, \cdot)$ are comonotone, and (5) implies

$$\int mdc = \sum_{k=1}^{K} r_k \left[c(S_k) - c(S_{k-1})\right] = \sum_{k=1}^{K} (r_k - r_{k+1})c(S_k). \;\; \diamond$$

Comonotone additivity (15) entails *positive homogeneity*, that is, $\int rmdc = r \int mdc$ when $r \geq 0$. Collecting preceding statements of this section, we get a nice characterization of convex capacities:

Proposition 3.4. (On convex capacities, Schmeidler (1986)) *Suppose an operator* $I : \mathbb{M} \to \mathbb{R}$ *is monotone* (14)*, comonotone additive* (15)*, and satisfies* $I(1_S) = 1$. *Then the following three statements are equivalent:*

- $I(\cdot) = \int \cdot dc$*, featuring a normalized convex capacity* c.
- *There exists a normalized convex capacity* c *such that*

$$I(m) = \min\{\mathbb{E}_P m \ | \text{probability measure P} \geq \text{c}\} \text{ for each m} \in \mathbb{M}.$$

- $I(m + \bar{m}) \geq I(m) + I(\bar{m})$ *for all* $m, \bar{m} \in \mathbb{M}$.

As already implicit in (15), it is convenient that several mappings be commonly monotone:

Proposition 3.5. (Choquet integral of a comonotone family on a finite state space) *Let* c *be a normalized capacity on a finite state space. Suppose all mappings* $m \in \mathcal{M} \subset \mathbb{M}$ *are strictly comonotone, meaning that all inequalities* (16) *are strict. Then there exists a probability measure* $P \geq c$ *such that*

$$\int m dc = \int m dP \text{ for all } m \in \mathcal{M}. \tag{17}$$

Proof. Fix any $m \in \mathcal{M}$ and order its range $m(S)$ as an increasing string $r_1 < \cdots < r_K$. Posit

$$P\{m = r_k\} := c\{m > r_{k-1}\} - c\{m > r_k\}.$$

This definition does not depend on the particular choice of m. (In particular, the cardinality of $m(S)$ is constant across $\mathcal{M}$.) Extend P, just defined, to become a probability measure on Σ. Now (10) tells that (17) holds. □

4 Integral Representation of Preferences

Basic to this paper is the assumption that utility be seen as money, hence as transferable. One advantage of this assumption was the prospect of finding price-generated core solutions via Lagrangian multipliers.

To secure an additional bonus, suppose that any payoff $\Pi(x)$, considered below, be fully determined from a corresponding monetary profile $m \in \mathbb{M}$. That is, with apologies for abuse of notation, suppose $\Pi(x) = \Pi(m(x))$. To elaborate on how and why this could happen, it is expedient to push the underlying, contingent claim $x \in \mathbb{X}$ temporarily back-stage, and – for a while – let merely $m = m(x)$ come to the fore. Then, what remains is a reduced criterion $\Pi(m)$.

Following Chateauneuf (1994), this section deals with the representation of criteria like $\Pi(m)$ as Choquet integrals. More generally, it considers the integral

representation of preferences over $\mathbb{M}$. As usual, by a preference is meant a reflexive, total and transitive order $\succsim$ on $\mathbb{M}$. The symbol $\sim$ signifies indifference, and $\succ$ means strict preference.

Clearly, if two payment profiles always swing in the same direction – that is, when the two are comonotone – neither can hedge the other. Against this backdrop, $\succsim$ is called *uncertainty averse* iff

$$m \sim m' \;\&\; (m' \text{ and } m'' \text{ are comonotone}) \Rightarrow m + m'' \succsim m' + m''.$$

Intuitively, while m'' provides no hedge for the comonotone m', it is apt to furnish some for m. Thereby, $m + m''$ becomes preferable to $m' + m''$.[13]

Plainly, uncertainty aversion isn't universal, but monotonicity or non-satiation seems nearly so. Reflecting on this fact, relation $\succsim$ is said to be *increasing* iff

$$m(s) \geq \bar{m}(s) + r \text{ for some constant } r > 0 \text{ and all } s \in S \;\Rightarrow m \succ \bar{m}.$$

On a more technical note, $\succsim$ is called *continuous* iff

$$\begin{cases} \bullet\; m^k \succsim \bar{m} \;\;\&\;\; m^k \downarrow m \text{ uniformly} \Rightarrow m \succsim \bar{m}, \text{ and} \\ \bullet\; m \succsim m^k \;\;\&\;\; m^k \uparrow \bar{m} \text{ uniformly} \Rightarrow m \succsim \bar{m}. \end{cases}$$

Admittedly, this continuity concept is rather demanding. The first bullet, for instance, requires that $\sup_{s\in S} \{m^k(s) - m(s)\} \searrow 0$, and $m^k(s) \geq m^{k+1}(s)$ for all s.

Now, if $\succsim$ indeed is uncertainty averse, increasing and continuous, it assumes a rather tractable form:

Theorem 4.1. (On integral representation of preferences, Chateauneuf (1994), Schmeidler (1986)) *For an increasing, continuous preference relation $\succsim$ on $\mathbb{M}$, the following two properties are equivalent:*

- *$\succsim$ is uncertainty averse;*
- *there exists a unique, normalized, convex capacity c such that*

$$m \succsim \bar{m} \Leftrightarrow \int m dc \geq \int \bar{m} dc.$$

In particular, since $m \sim I(m)1_S$, the integral $I(\cdot)$ associates to m its *certainty equivalent* $I(m) = \int m dc$.

Henceforth, assume that each *agent a has an uncertainty-averse, increasing, continuous preference relation $\succsim_a$ over $\mathbb{M}$*. Consequently, from Theorem 4.1, to any $a \in A$ is associated a unique normalized convex capacity c_a such that $m \succsim_a \bar{m} \Leftrightarrow \int m dc_a \geq \int \bar{m} dc_a$.

[13] Uncertainty aversion, as introduced here, bears resemblance to *variance aversion,* meaning that $m \succsim m + m''$ whenever $\mathbb{E}m'' = 0$ and $cov(m, m'') = 0$.

5 Explicit Core Solutions

The preceding section pushed contingent claims out of sight. It is time to call them on stage again – and be more specific about their nature and impact.

Henceforth, let $\mathbb{X}$ consist of all bounded, Σ-measurable mappings x from the state set S into the Euclidean commodity space $\mathcal{X}$. Further, suppose the monetary profile $m_a \in \mathbb{M}$, that accrues to agent a after choosing $x_a \in \mathbb{X}$, has the form

$$s \mapsto m_a(s) = \pi_a(x_a(s), s) =: \pi_a(x_a)(s). \tag{18}$$

Prominent in (18) is a state-dependent payoff function $\pi_a(\cdot, s) : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$. That bivariate function $\pi_a(\cdot, \cdot)$ is upper semicontinuous in the first variable and Σ-measurable in the second. The function value $-\infty$ indicates violation of implicit constraints, if any; see Example 5.1 below. It is tacitly assumed that agent a, in each state s, takes care to choose $x_a(s) \in \mathcal{X}$, if possible, so that $\pi_a(x_a(s), s) > -\infty$.

If $\pi_a(\cdot, s)$ does not depend on s, we speak of *state-independent payoff*. If moreover, π_a is concave, agent a exhibits commonplace *risk aversion*. *Uncertainty aversion,* as formulized here, corresponds to the presence of a convex capacity.[14]

Example 5.1. (*On individual payoff*) Suppose agent a, in state s, while owning the realized commodity bundle $\chi \in \mathcal{X}$, seeks to

$$\text{maximize } f_a(y_a, s) \text{ in the decision variable } y_a \in Y_a(s) \text{ s.t. } g_a(y_a, s) \geq \chi. \tag{19}$$

Here $\geq$ denotes a standard vector order in the commodity space $\mathcal{X}$, and $g_a(\cdot, s)$ maps the state-contingent "decisions space" $Y_a(s)$ into $\mathcal{X}$. If problem (19) is feasible, let $\pi_a(\chi, s)$ denote its optimal value; otherwise posit $\pi_a(\chi, s) = -\infty$. When $Y_a(s)$ is convex, and $f_a(\cdot, s)$, $g_a(\cdot, s)$ are both concave, $\pi_a(\cdot, s)$ becomes concave as well. $\diamond$

By hypothesis, agent a cares only about his monetary profile $s \mapsto m_a(s) = \pi_a(x_a(s), s)$. Moreover, his preference $\succsim_a$ over such profiles $m_a \in \mathbb{M}$ is presumed uncertainty averse, increasing and continuous. Hence, by Theorem 4.1, his criterion takes the form of *Choquet expected utility*, featuring a convex capacity c_a:

$$\Pi_a(x_a) := \int \pi_a(x_a) dc_a = \min \{\mathbb{E}_P \pi_a(x_a) \mid \text{probability measure } P \geq c_a)\}. \tag{20}$$

Format (20) generalizes the classical model of expected utility. It was arrived at somewhat indirectly here. For a direct axiomatization, see Gilboa (1987). As said,

[14] Both sorts of aversion affect portfolio choice; see Gollier (2006), Maenhout (2004). In general, uncertainty or ambiguity enters when a model must be specified. On such occasions, a prudent agent had better worry about making robust decisions. For studies of such issues in dynamic settings see Anderson et al. (2003), Cagetti et al. (2002).

the convexity of c_a mirrors some *pessimism* on the part of agent a. It also helps to identify or preserve some chief features; see Tallon (1998), Wakker (1990). Among those, the following are fairly immediate.

Proposition 5.1. (On concavity and monotonicity of criteria)

- *If the integrand $\pi_a(\chi, s)$ in (20) is concave (increasing) in $\chi \in X$, then $\Pi_a(\cdot)$ is also concave (respectively increasing).*
- *When each Π_a is concave (increasing), that property is transferred to the convoluted criterion $\Pi_A(\cdot)$ as well.*
- *Suppose each integrand $\pi_a(\chi, s)$ is concave in χ. Then, for any aggregate endowment e_A which yields finite overall payoff $\Pi_A(e_A)$, there exist probability measures $\bar{P}_a \geq c_a$, $a \in A$, such that*

$$\Pi_A(e_A) = \sup \left\{ \sum_{a\in A} \int \pi_a(x_a) d\bar{P}_a \;\middle|\; \sum_{a\in A} x_a = e_A \right\}.$$

Proof. Plainly, picking any $P \geq c_a$, the first implication holds for $\Pi_a(x, P) := \mathbb{E}_P \pi_a(x)$ hence for $\Pi_a(x) = \min\{\Pi_a(x, P) : P \geq c_a\}$. This takes care of the first bullet. The second is immediate.

For the third, note that $\{P_a \mid P_a \geq c_a\}$ is non-empty convex and compact.[15] Therefore, the lopsided min-max theorem in Aubin and Ekeland (1984) guarantees existence of a profile $a \mapsto \bar{P}_a \geq c_a$ such that

$$\begin{aligned}
\Pi_A(e_A) &= \sup \textstyle\sum_a \min \left\{ \int \pi_a(x_a) dP_a \mid P_a \geq c_a \right\} \text{ s.t. } \sum_a x_a = e_A \\
&= \min \sup \left\{ \textstyle\sum_a \int \pi_a(x_a) dP_a \;\middle|\; \sum_a x_a = e_A \right\} \text{ s.t. } P_a \geq c_a \text{ for each } a \\
&= \sup \left\{ \textstyle\sum_a \int \pi_a(x_a) d\bar{P}_a \;\middle|\; \sum_a x_a = e_A \right\}.
\end{aligned}$$

□

At this juncture, the most classical instance merits a brief mention:

Proposition 5.2. (Separability under agreed-upon risk) *Suppose there is a probability measure P such that $c_a = P$ for each a. Then, letting*

$$\pi_A(\chi, s) := \sup \left\{ \sum_{a\in A} \pi_a(\chi_a, s) \;\middle|\; \sum_{a\in A} \chi_a = \chi \right\}, \tag{21}$$

it holds that

$$\Pi_A(e_A) = \int \pi_A(e_A(s), s) dP(s).$$

□

[15] When S is finite, this is trivial. Otherwise, compactness refers to the w^*-topology.

Example 5.2. (*On coordinated programs*, Evstigneev and Flåm (2001), Flåm and Ermoliev (2008)) Referring to Example 5.1 and instance (19), $\pi_A(\chi, s)$, as defined by (21), equals the optimal value of the problem:

$$\text{maximize} \sum_{a\in A} f_a(y_a, s) \text{ s.t. } y_a \in Y_a(s) \ \& \ \sum_{a\in A} g_a(y_a, s) \geq \chi. \ \diamond$$

Clearly, the minimum taken in (20) preserves concavity of π_a, if any, but it tends to destroy smoothness. Worse and indeed difficult is the situation when agent a *likes uncertainty* – or is *optimistic*; see Tallon (1998), Wakker (1990) – in the sense that $c = c_a$ is concave, meaning that inequality (12) is reversed. Then Π_a might become non-concave, this causing numerous difficulties. In fact, there could be no core element at all, or no shadow price - hence no decentralized outcome.[16]

In core solution (3), one can reasonably hope that the endowment value λe_a has the integral format $\int_S \lambda(s)e_a(s)d\mu(s)$, featuring some probability measure μ on Σ alongside Σ-measurable linear mappings $\lambda(s) : \mathcal{X} \to \mathbb{R}$. This convenient format begs the question whether the "profit term" $\Pi_a^{(*)}(\lambda)$ in (3) also is separable. To explore this issue, let $\pi_a^{(*)}(\cdot, s)$ denote the conjugate of $\pi_a(\cdot, s)$.

Theorem 5.1. (State-separability of core solutions) *Suppose each integrand $\pi_a(\chi, s)$ is concave in χ. Also suppose each probability measure $P_a \geq c_a$ has a density f_a with respect to a measure μ on Σ for which the shadow price $\lambda x = \int_S \lambda(s)x(s)d\mu(s)$. Then there exists $\bar{P}_a \geq c_a$ with density $\bar{f}_a$ such that the core solution (3) takes the form*

$$\kappa_a(\lambda) = \int_S \left\{ \pi_a^{(*)}(\frac{\lambda(s)}{\bar{f}_a(s)}, s)\, \bar{f}_a(s) + \lambda(s)e_a(s) \right\} d\mu(s). \tag{22}$$

In case allocation, (x_a) is optimal, and λ is a shadow price, then

$$\lambda(s) \in \frac{\partial}{\partial\chi}\pi_a(x_a(s), s)\, \bar{f}_a(s) \ \textit{ almost surely for each } \ a. \tag{23}$$

Proof. As in the last bullet of Proposition 5.1 there exists a probability measure $\bar{P}_a \geq c_a$ with density $\bar{f}_a$ such that in the core solution (3) agent a receives cash $\kappa_a(\lambda) =$

$$\sup\left\{ \int \pi_a(x_a)d\bar{P}_a + \lambda(e_a - x_a) \mid x_a \in \mathbb{X} \right\}$$

$$= \sup\left\{ \int \left\{\pi_a(x_a(s), s)\, \bar{f}_a(s) + \lambda(s)\cdot[e_a(s) - x_a(s)]\right\} d\mu(s) \mid x_a \in \mathbb{X} \right\}$$

[16] And, of course, non-concave objectives would render the computation harder.

The last supremum commutes with the integral Rockafellar and Wets (1998). Consequently, using

$$\sup\left\{\pi_a(\chi,s)\,\bar{f}_a(s)-\lambda(s)\cdot\chi \;\middle|\; \chi\in\mathcal{X}\right\}=\pi_a^{(*)}\Big(\frac{\lambda(s)}{\bar{f}_a(s)},s\Big)\,\bar{f}_a(s),$$

inclusion (22) obtains. Specifically, when $\bar{f}_a(s)>0$, the last equation tells that $\lambda(s)$ must be a supergradient of $\pi_a(\cdot,s)\,\bar{f}_a(s)$ at the optimal choice $\chi=x_a(s)$. Otherwise, if $\bar{f}_a(s)=0$, then $\lambda(s)=0$. In either case, (23) is justified. □

Most applicable instances include finite state spaces, considered next. Serving well there is the counting measure μ with respect to which all probability measures are absolutely continuous.

6 Comonotone Allocations and Sunspots

To facilitate computation and display of explicit solutions, *this section takes S finite and let Σ comprise all subsets. Further, it uses $\mathbb{R}$ as a one-dimensional commodity space $\mathcal{X}$*. Accordingly, we posit now $\mathbb{M}=\mathbb{X}=\mathbb{R}^S$, and employ the ordinary dot product on this space.

Theorem 6.1. (Finite state space and explicit core solutions) *Suppose each $\pi_a(\chi,s)$ is concave in χ, and the convex capacity c_a is positive on each s. Then, for any shadow price $\lambda\in\mathbb{R}^S$ the cash payment scheme*

$$\kappa_a(\lambda):=\min\left\{\sum_{s\in S}\pi_a^{(*)}\Big(\frac{\lambda(s)}{P(s)},s\Big)P(s)\;\middle|\;P\geq c_a\right\}+\lambda\cdot e_a$$

belongs to the core.

Proof. Writing $c=c_a$ the assertion follows from

$$\sup_{x\in\mathbb{X}}\min_{P\geq c}\textstyle\sum_{s\in S}\min\{\pi_a(x(s),s)P(s)-\lambda(s)x(s)\}=$$

$$\min_{P\geq c}\sup_{x\in\mathbb{X}}\textstyle\sum_{s\in S}\{\pi_a(x(s),s)P(s)-\lambda(s)x(s)\}=$$

$$\min_{P\geq c}\textstyle\sum_{s\in S}\pi_a^{(*)}\big(\frac{\lambda(s)}{P(s)},s\big)P(s).$$

□

Core solutions are contracted *before* uncertainty resolves. Can those contracts be signed later, *after* the true s is unveiled? If payoff functions depend directly on s, or if probability assessments differ across agents, then agreements, normally underwritten upfront, can hardly be postponed. Put differently: it appears exceptional that

core imputations decided ex post, will implement those contracted ex ante.[17] After all, fire insurance cannot be had after the house has burned.

A well-noted exception is singled out next. It hinges upon formula (23) which dates, in the smooth case, back to Borch (1962), Wilson (1968). That formula sheds extra light on instances that feature state-independent payoff functions. If moreover, beliefs are somewhat aligned, then optimal allocations are apt to swing as does the aggregate; see Billot et al. (2000), Chateauneuf et al. (2000). To identify this feature, consider, for each $s \in S$, the ex post game in which coalition $C \subseteq A$ requires total enumeration no less than

$$\pi_C(\chi) := \sup\left\{ \sum_{a\in C} \pi_a(\chi_a) \;\middle|\; \sum_{a\in C} \chi_a = \chi \right\} \tag{24}$$

where $\chi = e_C(s) := \sum_{a\in C} e_a(s)$. Just as before, a cash payment scheme $k(s) = [k_a(s)] \in \mathbb{R}^A$ is declared in the *state-s core* iff it entails

$$\begin{cases} \textit{Pareto efficiency:} & \sum_{a\in A} k_a(s) = \pi_A(e_A(s)) \;\; \textit{and} \\ \textit{voluntary participation:} & \sum_{a\in C} k_a(s) \geq \pi_C(e_C(s)) \;\; \textit{for each coalition } C \subset A. \end{cases}$$

Together Theorems 2.1 and 2.2 tell that any state-contingent shadow price $\lambda(s) \in \partial\pi_A(e_A(s))$ generates a cash profile

$$a \mapsto k_a(s) := \kappa_a(\lambda(s)) := \pi_a^{(*)}(\lambda(s)) + \lambda(s) \cdot e_a(s)$$

that belongs to the state-s core. In these terms, it holds the following

Theorem 6.2. (Common uncertainty, comonotone allocation, and consistency) *Suppose each π_a is state-independent increasing, and there is a common capacity $c = c_a > 0$. Also suppose $\lambda(s) \in \partial\pi_A(e_A(s))$ for each $s \in S$. Then, any optimal allocation (x_a) yields, for each $s \in S$, an optimal allocation $[x_a(s)]$. These are comonotone and satisfy*

$$e_A(s) \leq e_A(s') \Rightarrow x_a(s) \leq x_a(s') \;\; \textit{for each} \;\; a. \tag{25}$$

Moreover, any probability measure $P \geq c$ such that $\int e_A dc = \int e_A dP$ satisfies

$$\Pi_a(x_a) = \int \pi_a(x_a(s))dP(s), \quad \Pi_a^{(*)}(\lambda) = \int \pi_a^{(*)}(\lambda(s))dP(s), \;\; \textit{and} \tag{26}$$

$$\kappa_a(\lambda) \;= \int \kappa_a(\lambda(s))dP(s) = \int \left[\pi_a^{(*)}(\lambda(s)) + \lambda(s) \cdot e_a(s)\right] dP(s) \; \textit{for each } a. \tag{27}$$

[17] The issue resembles that of *time consistency* Obstfeld and Rogoff (1996). But the setting here involves only *before* and *after* the state is unveiled - and not time proper.

In terms of such a measure P and the associated inner product

$$\langle \lambda, x \rangle := \int \lambda(s) \cdot x(s) dP(s)$$

it holds that $\lambda = [\lambda(s)] \in \partial \Pi_A(e_A)$. In short, λ supports an ex ante core solution that implements, for each state s, a corresponding ex post solution supported by $\lambda(s)$.

Proof. The argument will be coached so as to facilitate extension to infinite state spaces. By Theorem 2.2, the state-contingent shadow price $\lambda(s)$ satisfies $\lambda(s) \in \partial \hat{\pi}_A(e_a(s))$ where $\hat{\pi}_A(\cdot)$ is the smallest concave function $\geq \pi_A(\cdot)$, the latter being defined by (24). The concavity of $\hat{\pi}_A(\cdot)$ entails that $e_A(s) \leq e_A(s') \Rightarrow \lambda(s) \geq \lambda(s')$.

Similarly, looking at any agent a, it holds $\lambda(s) \in \partial \hat{\pi}_a(x_a(s))$ and $\lambda(s') \in \partial \hat{\pi}_a(x_a(s'))$ with $\lambda(s) \geq \lambda(s')$. Consequently, $x_a(s) \leq x_a(s')$. This takes care of (25) and the asserted comonotonicity. That property is further transferred to increasing payoffs:

$$x_a(s) \leq x_a(s') \Rightarrow \pi_a(x_a(s)) \leq \pi_a(x_a(s')) \text{ for each } a.$$

Because $\pi_a^{(*)}(\cdot)$ is decreasing, it holds $e_A(s) \leq e_A(s') \Rightarrow \pi_a^{(*)}(\lambda(s)) \leq \pi_a^{(*)}(\lambda(s'))$. The upshot is that all functions $\pi_a(x_a(s))$, $\pi_a^{(*)}(\lambda(s))$ for $a \in A$, are comonotone. Then any probability measure $P \geq c$ such that $\int e_A dc = \int e_A dP$ satisfies (26); see Chateauneuf et al. (2000).

Since P vanishes only on $\varnothing$, any linear $\lambda : \mathbb{X} \to \mathbb{R}$ may be represented on the form $\langle \lambda, x \rangle = \int \lambda(s) \cdot x(s) dP(s)$. Using that representation (27) obtains from

$$\Pi_A(e_A) = \sum_{a \in A} \int \left[\pi_a^{(*)}(\lambda(s)) + \lambda(s) e_a(s) \right] dP(s),$$

and $\lambda = [\lambda(s)] \in \partial I_A(e_A)$. □

State-independence isn't always realistic Karni (1993) but a weaker form merits separate mention. Suppose S comes as a product set $S_0 \times S_1$, but only component $s_0 \in S_0$ of any state $s = (s_0, s_1)$ matters. That is, suppose the aggregate endowment $e_A(s_0, s_1)$ and each individual payoff function $\pi_a(\chi, s_0, s_1)$ be totally unaffected by s_1. If so, component s_1 is called a "sunspot".

Can agents safely ignore sunspots? Under expected utility, they cannot, unless beliefs are common; see Cass and Shell (1983). Following Tallon Tallon (1998), given convex capacities (i.e. commonplace pessimism), it suffices for the irrelevance of sunspots to have some commonality in opinions about the likelihoods of such spots:

Proposition 6.2. (On irrelevance of sunspots) *Let $S = S_0 \times S_1$, but suppose all payoffs $\pi_a(\chi, s_0, s_1)$ be strictly concave in χ and independent of s_1. Further assume e_A is unaffected by s_1. Suppose there exists a probability measure $\bar{P} \geq$ all c_a such that for each $s_0 \in S_0$ its conditional distribution $\bar{P}(\cdot\,|s_0)$ over S_1 has full support. Then, each optimal allocation $[x_a(s_0, s_1)]$ is almost surely independent of s_1. That is, sunspots do not matter.*

Proof. Here it is worthwhile to ignore the standing assumption (of this section) that S be finite. Pick any probability distribution $\bar{P}$ that has the described properties. With allocation (x_a) optimal, define a new feasible allocation $(\bar{x}_a)$ as an average over sunspots, namely for any $s_0 \in S_0$ posit

$$\bar{x}_a(s_0) := \int x_a(s_0, s_1) d\bar{P}(s_1\,|s_0).$$

By concavity

$$\pi_a(\bar{x}_a(s_0), s_0) \geq \int \pi(x_a(s_0, s_1), s_0) d\bar{P}(s_1\,|s_0)$$

with strict inequality if $x_a(s_0, \cdot)$ is not $\bar{P}(\cdot\,|s_0)$-almost surely constant across S_1. Thus, unless $x_a(s_0, s_1)$ is not almost everywhere unaffected by s_1 for all a, it follows that

$$\Pi_A(e_A) \geq \sum_{a \in A} \Pi_a(\bar{x}_a) \geq \sum_{a \in A} \mathbb{E}_{\bar{P}} \pi_a(\bar{x}_a) > \sum_{a \in A} \mathbb{E}_{\bar{P}} \pi_a(x_a)$$

$$\geq \sum_{a \in A} \min\{\mathbb{E}_P \pi_a(x_a) : P \geq c_a\} = \sum_{a \in A} \Pi_a(x_a) = \Pi_A(e_A).$$

This contradiction concludes the argument. □

Equation (13) has driven the above results. With S finite as here, that equation may be relaxed a bit. To wit, in the same setting as Example 3.0 (10) implies

$$\int m dc \leq r_1 + (r_2 - r_1) P\{m > r_1\} + \cdots + (r_n - r_{n-1}) P\{m > r_{n-1}\} = \int m dP$$

for each probability measure $P \geq c$. Then, using the convention that $\min \emptyset = +\infty$, any normalized capacity c satisfies

$$\int m dc \leq \min\left\{ \int m dP \mid P \geq c \right\}.$$

Anyway, (20) helps to divorce different features of the agent's attitudes. On one side, the integrand π_a reflects his taste for income or wealth. On the other side,

when there are many probability measures $P \geq c_a$, uncertainty aversion prevails Schmeidler (1989). So, paraphrasing Yaari (1985), *diminishing marginal utility of money* and *uncertainty aversion* are no longer two colours of the same horse.

Acknowledgements Thanks are due to Finansmarkedsfondet for financial support and a referee for good and generous comments.

References

Anderson, E. W., Hansen, L. P., & Sargent, T. J. (2003). A quartet of semigrous for model specification, robustness, prices of risk, and model detection. *J European Economic Association*, *1*(1), 68–123.

Aubin, J. -P., & Ekeland, I. (1984). *Nonlinear Applied Analysis*. New York: Wiley.

Baton, B., & Lemaire, J. (1981). The core of a reinsurance market. *ASTIN Bulletin*, *12*, 57–71.

Billot, A., Chateauneuf, A., Gilboa, I., & Tallon, J. -M. (2000). Sharing beliefs: between agreeing and disagreeing. *Econometrica*, *68*, 685–694.

Borch, K. H. (1960). Reciprocal reinsurance treaties. *ASTIN Bulletin*, *1*, 171–191.

Borch, K. H. (1960). Reciprocal reinsurance treaties seen as a two-person cooperative game. *Scandinavian Actuarial Journal*, *43*, 29–58.

Borch, K. H. (1962). Equilibrium in a reinsurance market. *Econometrica, 30*, 424–444.

Cagetti, M., Hansen, L. P., Sargent, T., & Williams, N. (2002). Robustness and pricing under uncertain growth. *The Review of Financial Studies, 15*(2), 363–404.

Cass, D., & Shell, K. (1983). Do sunspots matter? *Journal of Political Economy, 91*, 193–227.

Cass, D., Chichilnisky, G., & Wu, H. -M. (1996). Individual risk and mutual insurance. *Econometrica, 64*, 333–341.

Chateauneuf, A. (1994). Modeling attitudes towards uncertainty and risk through the use of Choquet integral, *Annals of Operations Research, 52*, 3–20.

Chateauneuf, A., Dana, R.-A., & Tallon, J. -M. (2000). Optimal risk-sharing rules and equilibria with Choquet-expected-utility, *Journal of Mathematical Economics, 34*, 191–214.

Denneberg, D. (1994). *Non-additive measure and integral*. Dordrecht: Kluwer.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of Royal Statistical Society B, 30*, 205–247.

Dow, J., Ribeiro, S., & Werlang, C. (1992). Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica, 60*(1), 197–204.

Evstigneev, I. V., & Flåm, S. D. (2001). Sharing nonconvex cost. *Journal of Global Optimization, 20*, 3–4, 257–71.

Fishburn, P. C. (1988). *Nonlinear preference and utility theory*. Baltimore: The Johns Hopkins University Press.

Flåm, S. D., & Ermoliev, Y. (2008). Investment, uncertainty and production games. *Environment and Development Economics*.

Flåm, S. D., Owen, G., & Saboya, M. (2005). The not-quite non-atomic game: non-emptiness of the core in large production games. *Mathematical Social Sciences, 50*, 279–297.

Flåm, S. D., & Koutsougeras, L. (2010). Private information, transferable utility, and the core. *Economic Theory, 42*, 591–609.

Gilboa, I. (1987). Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics, 16*, 65–88.

Gilboa, I., & Schmeidler, D. (1994). Additive representations of non-additive measures and the Choquet integral. *Annals of Operations Research, 52*, 43–65.

Karni, E., & Schmeidler, D. (1991). Utility theory with uncertainty. In W. Hildenbrandt & H. Sonnenschein (Eds.), *Handbook of Mathematical Economics* (vol IV, Chap. 33). Amsterdam: North-Holland.

Gollier, C. (2006). Does ambiguity aversion reinforce risk aversion? Applications to portfolo choice and asset prices. Typescript.

Karni, E. (1993). Subjective expected utility theory with state-dependent preferences. *Journal of Economic Theory, 69*, 428–438.

Laurent, P. -J. (1972). *Approximation et optimisation*. Paris: Hermann.

Machina, M. (1987). Choice under uncertainty: problems solved and unsolved. *Economic Perspectives, 1*, 121–154.

Maenhout, P. J. (2004). Robust portfolio rules and asset pricing. *The Review of Financial Studies, 17*(4), 951–983.

Obstfeld, M., & Rogoff, K. (1996). *Foundations of international macroeconomics*. MIT.

Osborne, M. J., Rubinstein, A. (1994). *A course in game theory*. MIT.

Rockafellar, R. T., & Wets, J. -B. (1998). *Variational analysis*. Berlin: Springer.

Schmeidler, D. (1986). Integral representation without additivity. *Proc. Am. Math. Soc., 97*, 255–261.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*, 571–587.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Shapley, L. S. (1971). Cores of convex games. *Int. Journal of Game Theory, 1*, 12–26.

Tallon, J. -M. (1998). Do sunspots matter when agenst are Choquet-expected-utility maximizers? *Journal of Economic Dynamics and Control, 22*, 357–368.

Wakker, P. (1990). Characterizing optimism and pessimism directly through comonotonicity. *Journal of Economic Theory, 52*, 453–463.

Wang, S. S., Young, V. R. & Panjer, H. H. (1997). Axiomatic characterization of insurance prices. *Insurance: mathematics and economics, 21*, 173–183.

Wilson, R. (1968). The theory of syndicates. *Econometrica, 36*, 119–132.

Yaari, M. E. (1985). The dual theory of choice under risk. *Econometrica*, 95–115.

The Perpetual American Put Option for Jump-Diffusions

Knut K. Aase

Abstract We solve a specific optimal stopping problem with an infinite time horizon, when the state variable follows a jump-diffusion. The novelty of the paper is related to the inclusion of a jump component in this stochastic process. Under certain conditions, our solution can be interpreted as the price of an American perpetual put option. We characterize the continuation region when the underlying asset follows this type of stochastic process. Our basic solution is exact only when jump sizes cannot be negative. A methodology to deal with negative jump sizes is also demonstrated.

1 Introduction

We consider the problem to price an American put option when the underlying asset pays no dividends, and the time to expiration is infinite (the perpetual case). As shown by Samuelson (1965), this problem is equivalent to the pricing of an infinite-lived American call option that pays a continuous, proportional dividend rate.

The market value of the corresponding European perpetual put option is known to be zero, but as shown by Merton (1973a), the American counterpart converges to a strictly positive value. This demonstrates at least one situation where there is a difference between these two products in the case with no dividend payments from the underlying asset.

We analyze this contingent claim when the underlying asset has jumps in its price path. We start by solving the relevant optimal stopping problem for a fairly general class of jump-diffusions. Our method does not provide the overall solution when jumps can be negative, but it is possible to construct an approximation via Dynkin's formula also for that instance – as demonstrated in the present paper.

K.K. Aase
Department of Finance and Management Science, Norwegian School of Economics and Business Administration (NHH), Helleveien 30, 5045 Bergen, Norway
and
Centre of Mathematics for Applications (CMA), University of Oslo, Oslo, Norway
e-mail: knut.aase@nhh.no

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_28,
© Springer-Verlag Berlin Heidelberg 2010

The pricing of American options, while subject to intensive and extensive research during the last three decades, has found no closed form solution except in specific cases. A variety of numerical schemes has therefore been developed, as may be inferred from the (far from exhaustive) list of references. However, the open-ended nature of the setting – that is, the absence of maturity – simplifies the problem, just like in control of Markov chains.

The present paper adds to – and is inspired (and informed) by – the seminal Samuelson (1965) paper on warrant pricing, as well as Merton's paper (1973). The infinite horizon and the assumed time homogeneity makes one expect a time-invariant solution – a conjecture proved below, using the verification theorem.

The paper is organized as follows: Section 2 presents the model, Sect. 3 the pricing problem, and Sect. 4 gives the solution. Section 5 deals with risk adjustments, Sect. 6 considers negative jump sizes, and Sect. 7 concludes.

2 The Model

We start by establishing the dynamics of the assets in the model: There is an underlying probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ satisfying the usual conditions, where Ω is the set of states, $\mathcal{F}$ is the set of events, $\mathcal{F}_t$ is the set of events observable by time t, for any $t \geq 0$, and P is the given probability measure, governing the probabilities of events related to the stochastic price processes in the market. On this space is defined one (locally) riskless asset, thought as the evolution of a bank account with dynamics

$$d\beta_t = r\beta_t dt, \qquad \beta_0 = 1,$$

and one risky asset satisfying the following stochastic differential equation

$$dS_t = S_{t-}[\mu dt + \sigma dB_t + \alpha \int_Z \eta(z)\tilde{N}(dt, dz)], \qquad S_0 = x > 0. \tag{1}$$

Here B is a standard Brownian motion, $\tilde{N}(dt, dz) = N(dt, dz) - \nu(dz)dt$ is the compensated Poisson random measure, $\nu(dz)$ is the Lévy measure, and $N(t, A)$ is the number of jumps which occur before or at time t with sizes in the set $A \subseteq Z$. The process $N(t, Z)$ is called the Poisson random measure of the underlying Lévy process. When this process makes a jump of size z at time t, the price process makes a jump of size $S_{t-}\,\alpha\,\eta(z)$. The state $S = 0$ is absorbing. For the price to remain non-negative, assume $\alpha\eta(z) \geq -1$ for all values of z. We will usually choose $\eta(z) = z$ for all z, which implies that the integral is over the set $Z = (-1/\alpha, \infty)$. The Lévy measure $\nu(A) = E[N(1, A)]$ is in general a set function, where E is the expectation operator corresponding to the probability measure P. In our examples, we will by and large assume that this measure can be decomposed into $\nu(dz) = \lambda dF(z)$ where λ is the frequency of the jumps and F is the probability distribution function of the jump sizes. This gives us a finite Lévy measure, and the jump part becomes a compound Poisson process.

This latter simplification is not required to deal with the optimal stopping problem, which can in principle be solved for any Lévy measure ν, provided the relevant equations are well defined, subject to certain technical conditions to which we return later. The processes B and N are assumed independent.

The stochastic differential equation (1) can be solved using Itô's lemma, giving

$$S(t) = S(0)\exp\Big\{(\mu - \frac{1}{2}\sigma^2)t + \sigma B_t - \alpha\int_0^t\int_Z \eta(z)\nu(dz)ds + \int_0^t\int_Z \ln(1+\alpha\eta(z))N(ds,dz)\Big\}. \tag{2}$$

From (2) one sees why we required $\alpha\eta(z) \geq -1$ for all z. (Taking the logarithm of negative number makes no sense.) Solution (2) is sometimes labeled a "stochastic" exponential, in contrast to only an exponential process which would result if the price Y was instead given by $Y(t) = Y(0)\exp(X_t - \frac{1}{2}\sigma^2 t))$, where the accumulated return process X_t is given by the arithmetic process

$$X_t := \mu t + \sigma B_t + \alpha\int_0^t\int_R \eta(z)\tilde{N}(ds,dz). \tag{3}$$

Clearly the process Y can never reach zero in a finite amount of time if the jump term is reasonably well behaved[1], so there would be no particular lower bound for the term $\alpha\eta(z)$ in this case. We have chosen to work with stochastic exponential processes in this paper. There are several reasons why this is a more natural model in finance. On the practical side, bankruptcy can be modeled using S, so credit risk issues are more readily captured by this model. Also the instantaneous return $\frac{dS(t)}{S(t-)} = dX_t$, which equals ($\mu dt +$ "noise"), where μ is the rate of return, whereas for the price model Y we have that

$$\frac{dY(t)}{Y(t-)} = \big(\mu + \int_R (e^{\alpha\eta(z)} - 1 - \alpha\eta(z))\nu(dz)\big)\,dt + \sigma dB_t + \int_R (e^{\alpha\eta(z)} - 1)\tilde{N}(dt,dz),$$

which is in general different from dX_t, and as a consequence we do not have a simple interpretation of the rate of return in this model. [2]

3 The Optimal Stopping Problem

We want to solve the following problem:

$$\psi(x) = \sup_{\tau\geq 0} E^x\Big\{e^{-r\tau}(K - S_\tau)^+\Big\}, \tag{4}$$

[1] i.e., if it does not explode. The Brownian motion is known not to explode.

[2] If the exponential function inside the two different integrals can be approximated by the two first terms in its Taylor series expansion, which could be reasonable if the Lévy measure ν has short and light tails, then we have $\frac{dY(t)}{Y(t-)} \approx dX_t$.

where $K > 0$ is a fixed constant, the exercise price of the put option, τ is a stopping time, and the dynamics of the stock follows the jump-diffusion process explained above. By E^x we mean the conditional expectation operator given that $S(0) = x$, under the given probability measure P.

For this kind of dynamics, the financial model is in general not complete[3], so in our framework the option pricing problem may not have a unique solution, or any solution at all. There will normally be many risk-adjusted measures Q, and if it is not even clear that the pricing rule must be linear, none of these may be appropriate for pricing the option at hand. If there is a linear pricing rule, however, the pricing problem may in some cases be a variation of the solution to the above problem, since under any appropriate Q the price S follows a dynamic equation of the type (1), with r replacing the drift parameter μ, and possibly with a different Lévy measure $\nu^Q(dz)$, absolutely continuous with respect to $\nu(dz)$. Thus, we first focus our attention on the problem (4).

There are special cases where the financial problem has a unique solution; in particular, there are situations including jumps where the model either is, or can be made complete, in the latter case by simply adding a finite number of risky assets.

The stopping problem (4) has been considered by other authors from different perspectives. Mordecki (2002) finds formulas based on extending the theory of optimal stopping of random walks. Its usefulness hinges upon one's ability to compute the quantity $E(e^I)$, where $I = \inf_{0\le t\le\tau(r)} Z(t)$, and $\tau(r)$ is an exponential random variable with parameter $r > 0$, independent of $Z(t) = X_t - \frac{1}{2}\sigma^2 t$, and $\tau(0) = \infty$. No adjustments to risk were considered. See also Boyarchenko and Levendroskiĭ (2002).

In contrast, we base our development on the theory of integro-variational inequalities for optimal stopping. Although we do not obtain exact solutions in all situations considered, our procedure is well suited to many applications of the option pricing theory.

4 The Solution of the Optimal Stopping Problem

In this section, we present the solution to the optimal stopping problem (4) for jump-diffusions. Let $\mathcal{C}$ denote the continuation region, and let τ be the exercise time defined by $\tau = \inf\{t > 0; S(t) \notin \mathcal{C}\}$. We make the assumption that

$$S(\tau) \in \bar{\mathcal{C}} \qquad (\bar{\mathcal{C}} \text{ is the closure of } \mathcal{C}). \tag{5}$$

In other words, at the time of jump τ, the process is not allowed to jump totally out of the continuation region. It is this condition that causes problems with negative jumps, as will be apparent later.

[3] There is, in general, too high a degree of uncertainty compared to the number of underlying assets (here two) for the model to be dynamically complete. However, a pure jump model with deterministic jump sizes is known to be complete.

We now have the following result:

Theorem 1. *The solution* $\psi(x) := \psi(x;c)$ *of the optimal stopping problem is, under the assumptions (5), given by*

$$\psi(x) = \begin{cases} (K-c)\left(\frac{c}{x}\right)^{\gamma}, & \text{if } x \geq c; \\ (K-x), & \text{if } x < c, \end{cases} \tag{6}$$

where the continuation region $\mathcal{C}$ *is given by*

$$\mathcal{C} = \{(x,t) : x > c\},$$

and the trigger price c *is a constant. This constant is given by*

$$c = \frac{\gamma K}{\gamma + 1}, \tag{7}$$

where the constant γ *solves the following equation*

$$-r - \mu\gamma + \frac{1}{2}\sigma^2\gamma(\gamma+1) + \int_Z \{(1+\alpha\eta(z))^{-\gamma} - 1 + \alpha\gamma\eta(z)\}\nu(dz) = 0. \tag{8}$$

Proof. As with continuous processes, there is an associated optimal stopping theory also for discontinuous ones. For an exposition, see Øksendal and Sulem (2004), for example. In order to employ this theory, we need the characteristic operator, or generator $\bar{\mathcal{A}}$ of the process S. For any smooth function $f : R \to R$ not depending upon time, it is defined as

$$\bar{\mathcal{A}}f(x) = \lim_{t \to 0+} \frac{1}{t}\{E^x[f(S_t)] - f(x)\} \qquad \text{(if the limit exists)},$$

where $E^x[f(S_t)] = E[f(S_t^x)], S_0^x = x$. Thus $\bar{\mathcal{A}}$ represents the expected rate of return of continuing at $t = 0$. For a time-homogeneous problem this is the expected rate of continuing at any time $t > 0$ as well. For our price process and with this kind of time-homogeneous function f, the generator for a jump-diffusion takes the following form:

$$\begin{aligned} \bar{\mathcal{A}}f(x) = {} & x\mu\frac{df(x)}{dx} + \frac{1}{2}x^2\sigma^2\frac{d^2f(x)}{dx^2} \\ & + \int_Z \left\{ f(x + \alpha x\eta(z)) - f(x) - \alpha\frac{df(x)}{dx}x\eta(z) \right\} \nu(dz), \end{aligned}$$

where the last term stems from the jumps of the price process S. Since the objective function depends upon time via the discount factor, our problem can be classified

as a time-inhomogeneous one. The standard theory of optimal stopping, and in particular the verification theorem, is formulated for the time-homogeneous case, but augmenting the state space of S by one more state, namely time itself,

$$Z_t = \begin{pmatrix} s+t \\ S_t \end{pmatrix}; \qquad t \geq 0$$

transforms the problem into a time-homogeneous one in the variable Z. (When $t = 0$, the process $Z(0) = (s, x)$.) It is now convenient to reformulate our problem as follows: We seek the discounted value function $\varphi(s, x)$ defined by

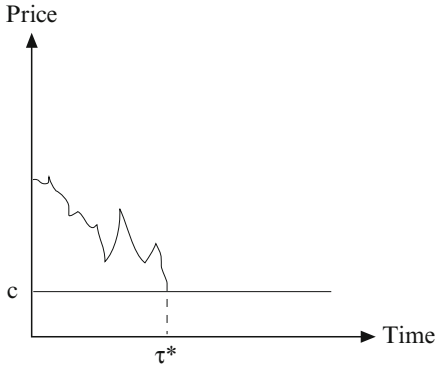
$$\varphi(s, x) := \sup_{\tau \geq 0} E^{(s,x)} \left\{ e^{-r(s+\tau)} (K - S_\tau)^+ \right\}. \tag{9}$$

The generator $\mathcal{A}$ of the process Z is given in terms of the generator $\bar{\mathcal{A}}$ of the process S by $\mathcal{A} f(s, x) = \bar{\mathcal{A}} f(s, x) + \frac{\partial}{\partial s} f(s, x)$, where $\bar{\mathcal{A}}$ works on the x-variable.

With a view toward the verification theorem – a version for jump-diffusion processes exists along the lines of the one for continuous processes – we now conjecture that the continuation region $\mathcal{C}$ has the following form

$$\mathcal{C} = \{(x, t) : x > c\},$$

where the trigger price c is some constant. The motivation for this is that for any time t the problem appears just the same, from a prospective perspective, implying that the trigger price $c(t)$ should not depend upon time. See Fig. 1. In the continuation region $\mathcal{C}$, the principle of optimal stopping requires $\mathcal{A}\varphi = 0$, or



Continuation region of the perpetual American put option

Fig. 1 Continuation Region

$$\frac{\partial \varphi}{\partial s} + \mu x \frac{\partial \varphi}{\partial x} + \frac{1}{2} x^2 \sigma^2 \frac{\partial^2 \varphi}{\partial x^2} + \int_Z \left\{ \varphi(s, x + \alpha x \eta(z)) - \varphi(s, x) - \alpha \frac{\partial \varphi}{\partial x} x \eta(z) \right\} \nu(dz) = 0.$$

This is a typical dynamic optimization criterion saying that it is not optimal to exercise so long as the expected rate of change of the value function is not strictly negative.

Furthermore, we conjecture that the function $\varphi(s, x) = e^{-rs}\psi(x)$. Substituting this form into the above equation allows us to cancel the common term e^{-rs}, and we are left with the equation

$$-r\psi(x) + \mu x \frac{\partial \psi(x)}{\partial x} + \frac{1}{2} x^2 \sigma^2 \frac{\partial^2 \psi(x)}{\partial x^2} + \int_Z \left\{ \psi(x + \alpha x \eta(z)) - \psi(x) - \alpha \frac{\partial \psi(x)}{\partial x} x \eta(z) \right\} \nu(dz) = 0 \tag{10}$$

for the unknown function ψ.

Thus, we were successful in removing time from the PDE, and reducing the equation to an ordinary integro-differential-difference equation.

The equation is valid for $c \leq x < \infty$. Given the trigger price c, let us denote the market value $\psi(x) := \psi(x; c)$. The relevant boundary conditions are then

$$\psi(\infty; c) = 0 \quad \forall c > 0 \tag{11}$$

$$\psi(c; c) = K - c \quad \text{(exercise)} \tag{12}$$

We finally conjecture a solution of the form $\psi(x) = a_1 x + a_2 x^{-\gamma}$ for some constants a_1, a_2, and γ. The boundary condition (11) implies that $a_1 = 0$, and the boundary condition (12) implies that $a_2 = (K - c)c^{\gamma}$. Thus, the conjectured form of the market value of the American put option is the following

$$\psi(x; c) = \begin{cases} (K - c)\left(\frac{c}{x}\right)^{\gamma}, & \text{if } x \geq c; \\ (K - x), & \text{if } x < c. \end{cases}$$

In order to determine the unknown constant γ, we insert this function in (10). This allows us to cancel the common term $x^{-\gamma}$, and we are left with the following nonlinear, algebraic equation for the determination of the constant γ:

$$-r - \mu\gamma + \frac{1}{2}\sigma^2 \gamma(\gamma + 1) + \int_Z \{(1 + \alpha\eta(z))^{-\gamma} - 1 + \alpha\gamma\eta(z)\} \nu(dz) = 0. \tag{13}$$

This is a well-defined equation in γ, and the fact that we have successfully been able to cancel out the variables x and s, is a strong indication that we actually have found the solution to our problem.

If this is correct, it only remains to find the trigger price c, and this we do by employing the "high contact" or "smooth pasting" condition (e.g., McKean 1965).

$$\frac{\partial \psi(c;c)}{\partial x}\Big|_{x=c} = -1.$$

This leads to the equation

$$(k-c)c^{\gamma}(-\gamma c^{-\gamma-1}) = -1,$$

which determines the trigger price c as

$$c = \frac{\gamma K}{\gamma + 1},$$

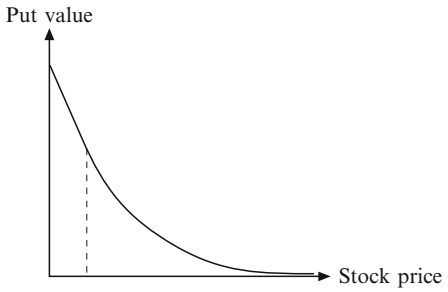
where γ solves (13) or (8). See Fig. 2.

We can now finally use the verification theorem of optimal stopping for jump-diffusions (see e.g. Øksendal and Sulem 2004) to prove that this is *the* solution to our problem. The main component of the verification theorem is the Dynkin formula, which states that

$$E^x\{\psi(S(\tau))\} = \psi(x) + E^x\left\{\int_0^{\tau} \mathcal{A}\psi(S(t))\,dt\right\}. \tag{14}$$

Here the requirement that $S(\tau) \in \bar{\mathcal{C}}$ comes into play, a sufficient condition for the theorem to hold. □

Remarks. 1. If we use the exponential pricing model Y defined in Sect. 2 instead of the stochastic exponential, where $Y(t) = Y(0)\exp(Z_t)$, $Z_t = (X_t - \frac{1}{2}\sigma^2 t)$ and the the accumulated return process X_t is given by the arithmetic process in (3),



Market value of a perpetual American put option as a function of stock price

Fig. 2 Perpetual American put option value

this problem also has a solution, the above method works, and the corresponding equation for γ is given by

$$-r-\gamma\left(\mu+\int_R(e^{\alpha\eta(z)}-1-\alpha\eta(z))\nu(dz)\right)+\frac{1}{2}\sigma^2\gamma(\gamma+1)+\int_R\left(e^{-\gamma\alpha\eta(z)}-1+\gamma(e^{\alpha\eta(z)}-1)\right)\nu(dz)=0. \tag{15}$$

2. By the verification theorem we get smooth pasting for free.
3. We may interpret the term $\left(\frac{x}{c}\right)^{-\gamma} I_{\{\tau(\omega)\in[t,t+dt)\}}(\omega)$ as the "state price" when $x \geq c$, where I indicates if exercise happens at time t or not: If exercise takes place at time t, then $(K-c)$ units are paid out at a price $(x/c)^{-\gamma}$ per unit when $x \geq c$, and $(K-x)$ units are paid at price 1 per unit if $x < c$. Hence, the term $(x/c)^{-\gamma}$ can be interpreted as an "average state price" when $x \geq c$.
4. The assumption (5) may seem restrictive at this point, as it basically rules out jump processes having negative jumps. The problem arises if exercise occurs at a jump time of S. When this jump is negative it may carry $S(\tau)$ inside the exercise region where the value function $\psi(\cdot)$ is linear according to (6), in which case Dynkin's formula does not apply, since the value function has another form inside the integral in (14), also illustrated in Fig. 2.

In Sect. 6, we indicate a direct method to solve the problem when jumps are negative, based on Dynkin's formula. We also demonstrate that the solution provided by Theorem 1 may still be a good approximation in this situation, especially when the current stock price x is away from the exercise boundary c.

5 Risk Adjustments

While the concept of an equivalent martingale measure is well known in the case of diffusion price processes with a finite time horizon $T < \infty$, the corresponding concept for jump price processes is less known. In addition, we have an infinite time horizon, in which case it is not true that the "risk neutral" probability measure Q is equivalent to the given probability measure P.

Suppose P and Q are two probability measures, and let $P_t := P|_{\mathcal{F}_t}$ and $Q_t := Q|_{\mathcal{F}_t}$ denote their restrictions to the information set $\mathcal{F}_t$. Then P_t and Q_t are equivalent for all t if and only if $\sigma^P = \sigma^Q$ and the Lévy measures ν^P and ν^Q are equivalent.

We now restrict attention to the pure jump case, where the diffusion matrix $\sigma = 0$. Let $\theta(s,z) \leq 1$ be a process such that

$$\xi(t) := \exp\left\{\int_0^t\int_Z \ln(1-\theta(s,z))N(ds,dz)+\int_0^t\int_Z\theta(s,z)\nu(dz)ds\right\} \tag{16}$$

exists for all t. Define Q_t by

$$dQ_t(\omega) = \xi(t)dP_t(\omega)$$

and assume that $E(\xi(t)) = 1$ for all t. Then there is a probability measure Q on $(\Omega, \mathcal{F})$ with the property that if we define the random measure $\tilde{N}^Q$ by

$$\tilde{N}^Q(dt, dz) := N(dt, dz) - (1 - \theta(t, z))\nu(dz)dt,$$

then

$$\int_0^t \int_Z \tilde{N}^Q(ds, dz) = \int_0^t \int_Z N(ds, dz) - \int_0^t \int_Z (1 - \theta(s, z))\nu(dz)ds$$

is a Q local martingale. Notice that $\tilde{N}^Q(dt, dz) = \tilde{N}(dt, dz) + \theta(t, z)\nu(dz)dt$.

This result can be used to prove the following version of Girsanov's theorem for jump processes:

Theorem 2. *Let S_t be a 1-dimensional price process of the form*

$$dS_t = S_{t-}[\mu dt + \alpha \int_Z \eta(z)\tilde{N}(dt, dz)].$$

Assume there exists a function $\theta(z) \leq 1$ such that

$$\alpha \int_Z \eta(z)\theta(z)\nu(dz) = \mu \qquad a.s. \tag{17}$$

and such that the corresponding process ξ_t given in (16) (with $\theta(s, z) \equiv \theta(z)$ for all s) exists, with $E(\xi_t) = 1$ for all t. Consider a measure Q such that $dQ_t = \xi(t)dP_t$ for all t. Then Q is a local martingale measure for S.

Proof. By the above-cited result and the equality (17) we have that

$$\begin{aligned} dS_t &= S_{t-}\left[\mu dt + \alpha \int_Z \eta(z)\tilde{N}(dt, dz)\right] \\ &= S_{t-}\left[\mu dt + \alpha \int_Z \eta(z)\{\tilde{N}^Q(dt, dz) - \theta(z)\nu(dz)dt\}\right] \\ &= S_{t-}\left[\alpha \int_Z \eta(z)\tilde{N}^Q(dt, dz)\right], \end{aligned}$$

which is a local Q-martingale. □

We will call Q a risk-adjusted probability measure, and θ the market price of risk (when we use the bank account as a numeraire). The above results can be extended to a system of n-dimensional price processes, see example, Øksendal and Sulem (2004) for results on a finite time horizon; Sato (1999), Chan (1999), Jacod and

Shiryaev (2002) and Cont and Tankov (2000) for general results; and Huang and Pagès (1992) or Revuz and Yor (1991) for results on the infinite time horizon.

Recall that the computation of the price of an American option must take place under a risk adjusted, local martingale measure Q in order to avoid arbitrage possibilities. Under any such measure Q, all the assets in the model must have the same rate of return, equal to the short-term interest rate r. Thus we should replace the term μ by r in (10). However, this may not be the only adjustment required when jumps are present. Typically another, but equivalent, Lévy measure $\nu^Q(dz)$ will appear instead of $\nu(dz)$ in (10). The details along this line will be considered elsewhere, together with applications.

6 Solution When Jumps Are Negative

In this section, we demonstrate a direct method based on Dynkin's formula to deal with the case of negative jump sizes. The problem with negative jumps stems from the change in the nature of the value function $\psi(x)$ when a jump brings the price process from the continuation region beyond the stopping boundary c and into the interior of the stopping region.

Our idea is simple, and best illustrated when jumps are discrete. Let us focus on one possible, negative jump size z_0. We divide the continuation region into disjoint parts, where we do know the nature of the value function $\psi(x)$ after a downward jump. Starting with the region $\mathcal{C}_1$ from which a jump brings the process into the stopping region where $\psi(x) = K - x$, we determine $\psi(x) = \psi_1(x)$ in $\mathcal{C}_1$. Next we consider the region $\mathcal{C}_2$ from which a jump brings the price into the region $\mathcal{C}_1$, where $\psi_1(x)$ has just been determined, and so on.

To this end, consider the pure jump model with negative jumps only

$$dS_t = S_{t-}\left[\mu dt + \int_{-1}^{\infty} z\tilde{N}(dt, dz)\right] \tag{18}$$

where the Lévy measure $\nu(dz) = \lambda\delta_{\{z_0\}}(z)dz$, and $z_0 < 0$ ($\alpha = 1$). Define the operator $\mathcal{A}$ by

$$\mathcal{A}\psi(x) = -r\psi(x) + \mu x\psi'(x) + \int_{-1}^{\infty} (\psi(x + zx) - \psi(x) - \psi'(x)z_0x)\nu(dz).$$

We want to find a constant $c \in (0, K)$ and a function ψ on $(0, \infty)$ such that ψ is continuous in $(0, \infty)$ and (a) $\psi(x) = K - x$ for $0 < x \leq c$ and (b) $\mathcal{A}\psi(x) = 0$ for $x > c$. We construct ψ on (c, ∞) by induction:

Case 1: $x \in \mathcal{C}_1 := (c, c/(1 + z_0))$. Then $x(1 + z_0) < c$ and therefore S jumps from $\mathcal{C}_1$ down to $(0, c)$ if it jumps, where ψ is given by (a). Thus condition (b) becomes

$$\mathcal{A}\psi(x) = -r\psi(x) + \mu x\psi'(x) + [K - x(1 + z_0) - \psi(x) - \psi'(x)z_0x]\lambda = 0$$

for $x \in \mathcal{C}_1$. This leads to the following standard first order in-homogeneous ODE

$$\psi'(x) + G(x)\psi(x) = H_1(x)$$

where $G(x) = -\frac{r+\lambda}{(\mu - z_0\lambda)x}$ and $H_1(x) = -\frac{[K - x(1+z_0)]\lambda}{(\mu - z_0\lambda)x}$. The solution, denoted $\psi_1(x)$ in $\mathcal{C}_1$, is

$$\psi_1(x) = e^{-\int_c^x G(v)dv}\left[\int_c^x e^{\int_c^v G(u)du} H_1(v)dv + k_1\right]. \tag{19}$$

By continuity of the value function, we determine the integrating constant k_1 by $\psi_1(c) = K - c$, implying that $k_1 = K - c$.

Case 2: $x \in \mathcal{C}_2 := (c/(1+z_0), c/(1+z_0)^2)$. Then $x(1+z_0) < c/(1+z_0)$ and therefore S jumps from $\mathcal{C}_2$ down to $\mathcal{C}_1$ if it jumps, where ψ is given by $\psi_1(\cdot)$ just determined. Thus condition (b) becomes

$$\mathcal{A}\psi(x) = -r\psi(x) + \mu x\psi'(x) + [\psi_1(x(1+z_0)) - \psi(x) - \psi'(x)z_0 x]\lambda = 0$$

for $x \in \mathcal{C}_2$. This leads to the same kind of ODE as above

$$\psi'(x) + G(x)\psi(x) = H_2(x)$$

where $G(x) = -\frac{r+\lambda}{(\mu - z_0\lambda)x}$ and $H_2(x) = -\frac{\psi_1(x(1+z_0))\lambda}{(\mu - z_0\lambda)x}$. The solution, denoted $\psi_2(x)$ in $\mathcal{C}_2$, is

$$\psi_2(x) = e^{-\int_{c/(1+z_0)}^x G(v)dv}\left[\int_{c/(1+z_0)}^x e^{\int_{c/(1+z_0)}^v G(u)du} H_2(v)dv + k_2\right]. \tag{20}$$

By continuity of the value function we determine the integrating constant k_2 by $\psi_1(c/(1+z_0)) = \psi_2(c/(1+z_0))$. Thus $k_2 = \psi_1(c/(1+z_0))$, where $\psi_1(\cdot)$ is given above. This determines the value function ψ on $\mathcal{C}_2$.

Next we define $\mathcal{C}_3 = (c/(1+z_0)^2, c/(1+z-0)^3)$ and proceed as above to determine ψ on $\mathcal{C}_3$ etc. We summarize as

Theorem 3. *The solution of the optimal stopping problem*

$$\phi(s,x) = \sup_{\tau \geq 0} E^{s,x}\left\{e^{-r(s+\tau)}(K - S_\tau)^+\right\},$$

with S_t given by (18) has the form $\phi(s,x) = e^{-rs}\psi(x)$ where $\psi(x)$ is given inductively by the above procedure. In particular we have that

$$\psi(x) = \begin{cases} K - x, & \text{for } 0 < x \leq c; \\ \psi_1(x) & \text{for } x \in \mathcal{C}_1; \\ \psi_2(x) & \text{for } x \in \mathcal{C}_2; \end{cases}$$

and $\psi(x) = \psi_n(x)$ *for* $x \in \mathcal{C}_n$, $n = 3, 4, \cdots$, *where* $\psi_1(x)$ *is given in (19),* $\psi_2(x)$ *is given in (20), etc.*

Since we here have a first order ODE, it is not a natural requirement that the first derivative of the value function $\psi'(x)$ is continuous in the patching point $x = c$. It is true that the function itself is continuous there, a requirement we have already used. Thus we may seem to be lacking a criterion to determine the trigger price c.

The solution $\psi(x)$ above is the value of an American perpetual put option if we adjust for risk, that is, when $\mu = r$ and λ is interpreted as the risk adjusted frequency under Q. If we consider the requirement $\psi'(c) = -1$, we only get a solution for c if $\mu \neq r$, and hence this trigger value does not correspond to the solution of the American put problem. One could perhaps conjecture that requiring the function $\psi(x)$ to be C^1 in the point $c/(1 + z_0)$ would provide the "missing" equation, but this turns out to yield a tautology, that is, $\psi_2'(c/(1 - z_0)) = \psi_1'(c/(1 + z_0))$ is automatically satisfied by the solution provided above and thus does not give anything new. The value of c must in fact be determined in the other end, namely by requiring that $\psi(x)$ approaches zero as $x \to \infty$.

Reexamining the exact procedure above, notice that if we approximate the linear function $(K - x(1 + z_0))$ by the curved one $\psi(x(1 + z_0))$ in the term dictating the inhomogeneous part of the first order ODE, we would obtain the solution given in Theorem 1. The effect of this perturbation will be more and more diluted as x increases. This we can see by comparing ψ_1 to ψ_2, where the linear term in the numerator of H_1 has already been replaced by a non-linear one in the numerator of H_2. Thus, we conjecture that for reasonably large values of the spot price x of the underlying asset, the solution obtained using Theorem 1 is a good approximation.

7 Conclusions

In the paper, we solved an optimal stopping problem with an infinite time horizon, when the state variable follows a jump-diffusion. Under certain conditions, explained in the paper, our solution can be interpreted as the price of an American perpetual put option, when the underlying asset follows this type of process.

Our basic solution is exact only when jump sizes cannot be negative. In the paper, we also investigate a solution technique approximation appropriate also for negative jumps.

The analysis in the paper has been theoretical in scope. However, the theory has several interesting applications. Just to illustrate, in a companion paper (Aase 2005) we have developed applications of the results on risk adjustments of jump-diffusions that may be used to shed some light on the equity premium puzzle, as well as the

risk-free rate puzzle. One key observation is that the probability distribution under the risk adjusted measure depends on the risk premium when jumps are present, while this is not the case for the standard, continuous version of the asset pricing model. This difference may be utilized, for example, to find risk premiums.

We then find the equity premium by investigating what risk premium it takes for the representative agent to just be indifferent to holding a long put on the stock market index. On the basis of the US stock market data of the last century and the analysis of this paper, this enables us to calibrate numbers. For example, when the risk free rate is 1%, the equilibrium premium on equity is found to be 2.5%, while the current estimate is 6% – the very cause of the equity premium puzzle. We conjecture that for the next 100 years we are not likely to experience this much reward for risk bearing.

References

Aase, K. K. (2008). "On the consistency of the Lucas pricing formula". *Mathematical Finance*, *18*(2), 293–303.

Aase, K. K. (2005). *"Using option pricing theory to infer about equity premiums"*. W.p., Anderson Graduate School of Management, UCLA.

Aase, K. K. (2002). "Equilibrium pricing in the presence of cumulative dividends following a diffusion". *Mathematical Finance*, *12*(3), 173–198.

Aase, K. K. (1999). "An equilibrium model of catastrophe insurance futures and spreads". *The geneva papers on risk and insurance theory*, *24*, 69–96.

Aase, K. K., Øksendal, B., & Ubøe, J. (2001). "Using the Donsker delta function to compute hedging strategies". *Potential Analysis*, *14*, 351–374.

Aase, K. K., Øksendal, B., Ubøe, J., & Privault, N. (2000). "White noise generalizations of the Clark-Haussmann-Ocone theorem with applications to mathematical finance". *Finance and Stochastics*, *4*(4), 465–496.

Boyarchenko, S. I., & Levendorskiĭ, S. Z. (2002). Non-Gaussian Merton-Black-Scholes theory. *Advanced series on statistics science and applied probability*, vol. 9. River Edge, NJ: World Scientific.

Chan, T. (1999). "Pricing contingent claims on stocks driven by Lévy processes". *Annals of Applied Probabilty*, *9*, 504–528.

Chimwanda, C. (2004). *The pricing of a perpetual American put option on an underlying driven by a jump diffusion*. Zimbabwe: University of Bulawayo (Preprint).

Cont, R., & Tankov, P. (2000). "*Financial modelling with jump processes*". Boca Raton, London, New York, Washington, DC: Chapman and Hall/CRC.

Huang, C. -F., & Pagès, H. (1992). "Optimal consumption and portfolio policies with an infinite horizon: existence and convergence". *Annals of Applied Probability*, *2*, 36–64.

Jacod, J., & Shiryaev, A. N. (2002). *Limit Theorems for Stochastic Processes* (2nd ed.). Berlin: Springer.

Lucas, R. (1978). "Asset prices in an exchange economy". *Econometrica*, *46*, 1429–1445.

McDonald, R., & Siegel, D. (1986). "The value of waiting to investment". *Quarterly Journal of Economics*, 707–727.

McKean, H. (1965). Appendix: "Free boundary problem for the heat equation arising from a problem in Mathematical Economics". *Industrial Management Review*, *6*, 32–39.

Merton, R. (1976). "Option pricing when the underlying stock returns are discontinuous". *Journal of Financial Economics*, *3*, 125–144.

Merton, R. C. (1973a). "Theory of rational option pricing". *Bell Journal of Economics and Management Science*, 141–183.

Merton, R. C. (1973b). "An intertemporal capital asset pricing model". *Econometrica*, *41*(5), 867–887.

Mordecki, E. (2002). "Optimal stopping and perpetual options for Lévy processes". *Finance Stochast*, *6*, 473–493.

Øksendal, B., & Sulem, A. (2004). *Applied stochastic control of jump diffusions*. Berlin, Heidelberg, New York: Springer.

Revuz, D., & Yor, M. (1991). *Continuous martingales and Brownian motion*. New York: Springer.

Samuelson, P. A. (1965). "Rational theory of warrant pricing". *Industrial Management Review*, *6*, 13–39.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge, U.K.: Cambridge University Press.

Discrete Event Simulation in the Study of Energy, Natural Resources and the Environment

Ingolf Ståhl

Abstract The development of computer technology has made discrete event simulation (DES) an increasingly attractive method. This chapter starts with a brief survey of the important uses of DES within the Energy, Natural Resource and Environmental Economics area. The chapter then describes three examples of how the work relating to this area using DES has done, namely: (1) models of project management, where simulation models allow for more realistic assumptions of time distributions and of limited resources than standard PERT (Program Evaluation Review Technique) and CPM (Critical Path Method) methods; (2) a bidding situation for oil resources, characterized by asymmetric information; (3) a small game dealing with duopolies producing and selling homogenous goods, such as oil or coal, but where demand is stochastic.

1 Introduction

The main purpose of this chapter is to show the role that discrete event simulation (DES) can play in a study program of Energy, Natural Resources and the Environment, in particular to give ideas for student projects in a simulation course within such a study program. I have no intention of going into the details of simulation programming, but I hope that I, by giving three concrete examples of simulation programs in this area, can convey the flavor of what small DES programs can do in this area.

I have had the privilege of running a course Simulation of Business Processes in cooperation with Kurt Jörnsten for a decade at NHH in Bergen. In this course, the students have first learnt the mechanics of a simulation software package, then some general features of the simulation process, e.g., input and output analysis, verification, and validation, and then they have proceeded to do a project on their

I. Ståhl
Center for Economic Statistics, Stockholm School of Economics, 11383 Stockholm, Sweden
e-mail: ingolf.stahl@hhs.se

E. Bjørndal et al. (eds.), *Energy, Natural Resources and Environmental Economics*, Energy Systems, DOI 10.1007/978-3-642-12067-1_29,
© Springer-Verlag Berlin Heidelberg 2010

own, dealing with a real system, generally in groups of two. The students have learnt much from doing these projects, and there are many potential problems within the area of Energy, Natural Resources and the Environment that would be suitable for such student projects.

With simulation, we refer to experiments with a computer model of a real system, where experiments refer to the systematic change in input variables, decisions, study the effects on output variables, results, draw conclusions to facilitate decisions. Here, we focus on DES dealing with stochastic systems, where the simulation model has to be run many times. DES plays an increasingly important role in business and has become the most important method in Management Science. The main reason for this is the extremely rapid development of computer technology. A computer for a given price has, and will probably within the nearest future, double in speed and capacity roughly every 18–24 months. This has implied that computers, for a given price, during the last two decades have increased in power more than a thousand times. This implies that if simulation earlier, because of high computing costs and time requirements, was regarded as a method of last resort, it is now by many regarded as the first alternative to try. Another factor contributing to this increased importance of simulation is the strong fall in prices for simulation software. Because of this, DES has come to replace many analytical/optimization parts of Management Science, e.g., queuing theory, inventory theory, PERT/CPM, and decision theory. Furthermore, with the increasing speed of computers and the development of new methods for optimization (e.g., tabu search, genetic algorithms, simulated annealing) simulation allowing for more realistic functions is being increasingly used also for optimization.

We have in our simulation course at NHH used a special simulation package, WebGPSS, a streamlined version of the General Purpose Simulation System (GPSS). GPSS, originally an IBM product, has been seen over a period of time as the most used special simulation language. WebGPSS is built on the experience from teaching GPSS to over 8,000 students (Born and Ståhl 2007). It is easy to learn, has a very easy-to-use graphical interface and is available also on the Web (www.webgpss.com).

In the literature, there are a great many examples on simulation within the area of Energy, Natural Resource and the Environment. One of the very first sub-areas where simulation was used was *mining*. The first mining simulation was done already in 1961, i.e., at the time of start of simulation. Sturgul et al. (2001) provide a survey of the use of simulation in mining with five different case studies of simulation projects in mining, all done in GPSS, the most popular software for simulation in mining. A simulation project on strip mining of oil sands (Shi and AbouRizk 1994) provides a link to another important sub-area, namely *construction* for environmental policy. In these sub-areas, we have several reports, which like many of the reports to be discussed here, have been presented at the WSC (the Winter Simulation Conference; www.wintersim.org), the most important annual conference within DES, with papers after 1996 available on the Web. Here, there are simulation reports on waste management for construction projects (Chandrakanthi et al. 2002), on the construction of a 6 km sewer collector (Halpin

and Martinez 1999) and on the construction of a dam embankment (Ioannou 1999). There are also reports that deal more generally with environmental simulation, like Kauffmann (1998).

More specifically there are reports on *energy*, like on energy market processes (Kapron and Kapron 2007), on the economic assessment of energy systems (Mallor et al. 2007), focusing on random events disturbing the supply of energy. Going down to specific sources of energy sources, there are reports dealing with *oil*. Examples deal with a web-based simulation game of an oil supply chain (Raychaudhuri 2007) and the risks when developing an oil field (Jacinto 2002). We have also had a student project at NHH, Improving Resource Utilization at a Bergen Oilfield Services Company, in 2002. There are also projects on *natural gas* (e.g., Conner et al. 1993). There are many reports on simulation of *electricity*, some more general, e.g., one comparing coal, nuclear and natural gas for the generation of electricity for households (Hamblin and Ratchford 2002). Other reports are on game theory models for electricity market simulations (Bompard and Abrate 2007), on simulations for mitigating risk in electricity generation (Brady 2001), and on simulation for improving the supply continuity in electric power systems (Nolan et al. 1993). Many reports deal with *nuclear* generation of electricity, e.g., on simulation for determining nuclear power plant operating crew size (Laughery et al. 1996) and on training nuclear security crews (Sanders and Lake 2005). Finally, it should be mentioned that I have had a student simulation project on a Swedish *bio-fuel* energy plant. Simulation for construction of wind mills should also be suitable as student projects.

2 Project Time Planning Simulation

We noted above that construction of environmental projects has been an important simulation topic. Construction projects are characterized as consisting of a number of tasks. Some tasks can be done in parallel with other tasks. Certain tasks cannot start until work on other tasks is complete. The time required for a given task is generally not fixed, but rather stochastic. Of particular interest is the time that the whole project will take. With random activity times, one wants to determine the probabilities for various total project times. Traditional methods for dealing with project time planning include PERT and CPM. Some of these methods require a fixed time for each activity, and some allow stochastic times, but are often limited to the Beta distribution. They do not always give correct estimates of the distribution of total project time, and those that do are generally limited to estimating the time variation only on the critical path, the longest time path through the project network. Lastly, resource limitations for the tasks are generally not considered explicitly.

DES solutions to project time planning do not have any of these limitations. To give some idea about how DES can be applied to construction problems, we give a simple example, regarding the construction of houses, but representative of any construction problem. [This example is built on Born and Ståhl (2008).] We assume that a contractor is planning to build a total of 100 houses. The contractor wants to

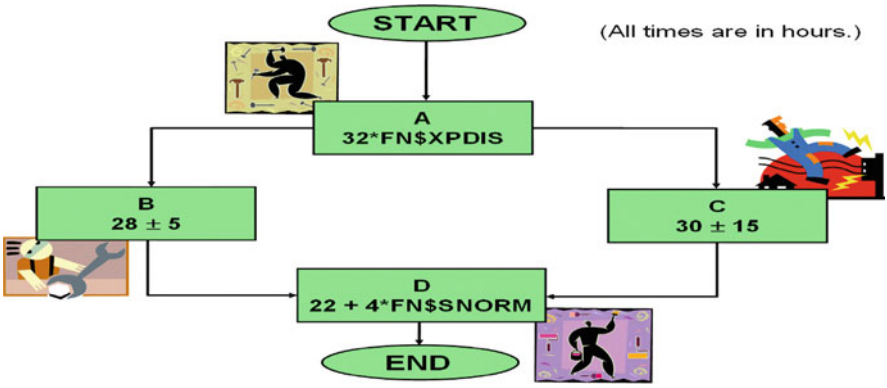


Fig. 1 Tasks in building a house

determine the total time required to build all the houses, as well as a distribution of the construction times for individual houses. The construction of a house involves four tasks: Process A is building the frame of the house; Processes B and C are plumbing and electrical wiring, respectively, which can begin only after the frame has been built, but can be done concurrently with each other; and Process D, which is painting, can be done only after the plumbing and electrical wiring tasks have been completed. Figure 1 provides a view of building a house and indicates the nature of the time distributions required for each of the four tasks. Building the frame is exponentially distributed with a mean of 32 h. Plumbing and painting are uniformly distributed with distributions of 28 ± 5 and 30 ± 15 h. Lastly, painting is normally distributed with a mean of 22 h and a standard deviation of 4 h. The WebGPSS built-in distribution fn$snorm has a mean 0 and standard deviation 1. For each of the processes shown in Fig. 1, there is only one worker available. When a worker has completed his part of the house, he can, if possible, start doing the same kind of work on the next house.

The GPSS program can be illustrated by the following GPSS block diagram (Fig. 2).

The GENERATE block creates 100 transactions = houses to be built. Construction of a house begins when it can seize the carpenter (aproc), building the frame of the house. The ARRIVE block marks the moment that the house construction begins. In the block ADVANCE 32*fn$xpdis, the carpenter builds the frame, and is then freed by the RELEASE block. The SPLIT block creates a copy of the transaction that goes to the electrician (cproc), while the original goes straight through the SPLIT to the plumber (bproc). When the plumber and electrician have completed their tasks and are freed by their corresponding RELEASE blocks, the two transactions, referring to the same house, are merged into one transaction after both of them reach the ASSEMBLE 2 block. The merged transaction then attempts to get service from the painter (dproc). When the painter is freed via its RELEASE block, the DEPART block measures the time used to construct the house. The transaction finally enters

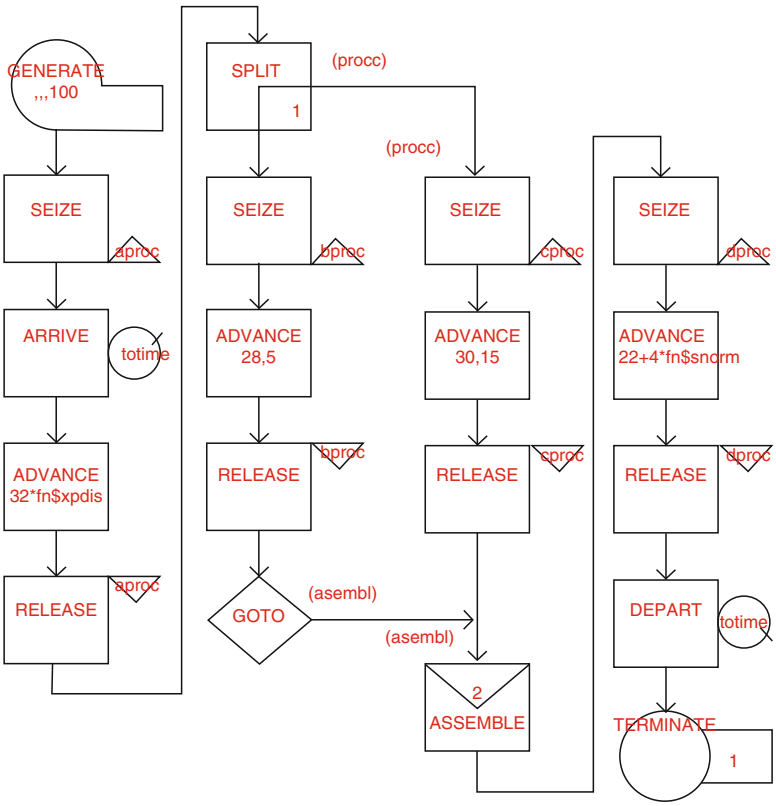


Fig. 2 Block diagram for the house construction project

the block TERMINATE 1 where the transaction leaves the system and removes one token. If the simulation starts with 100 tokens, the simulation will stop when the last token is removed, i.e., when the 100th house has been completed. A queue table is defined for the AD set *totime*. When the simulation is run, the resulting histogram provides a distribution of the times required to build the houses, as shown in Fig. 3.

We see great variations in the time required to construct a house. A majority of the 100 houses take between 100 and 200 h to build. Almost a third of the houses take between 200 and 500 h, and one house even requires between 500 and 600 h. To see how much of this variation can be due to randomness we replace the random times in the four advance blocks by their deterministic values, We then find that each house requires exactly 84 h to build. These 84 h correspond to the sum of the times along the critical path (longest time path in a PERT diagram), consisting of the processes A(32), C(30), and D(22). The differences between the stochastic and deterministic results are hence surprisingly large. This shows that stochastic time variations must be taken into consideration if one is interested in obtaining realistic and useful results.

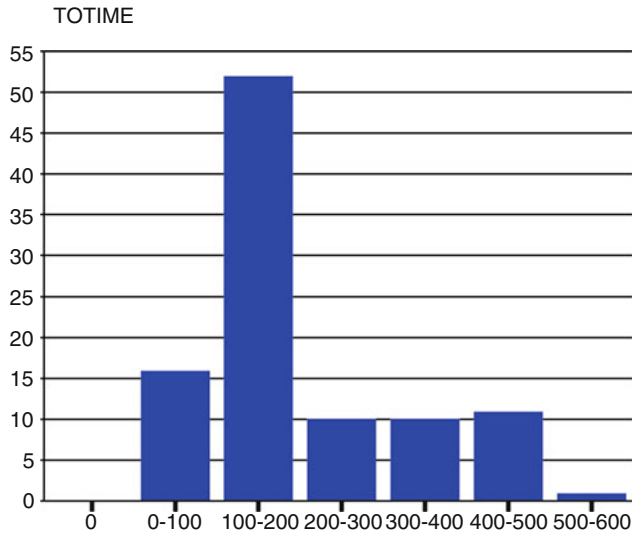


Fig. 3 Distribution of times to build houses

3 A Bidding Example

Samuelson and Bazerman (1985) provided a bidding problem in the energy sector, for which DES provides a powerful and illustrative solution. The problem is as follows: Company A is considering acquiring company T by means of a tender offer in cash for 100% of T's shares. The value of T depends on the outcome of an oil exploration project that T is currently undertaking. If the exploration fails completely, T will be worth \$0 per share, but in the case of a complete success, a share of T under current management will be worth \$100. All values between \$0 and 100 are equally likely. Regardless of the outcome, T will be worth 50% more under A's management than under its current management. The price of the offer to T must be determined before A knows the outcome of the drilling, but T will know the outcome when deciding whether or not to accept A's offer. T is expected to accept any offer from A that is greater than the per share value of T under its current management. What price should A offer?

A typical student's view is that the expected value of T to its current owner is \$50 and that is worth 50% more to A. Hence, A should bid in the interval of \$50–75. To demonstrate that this is wrong, we simulate the expected value for A for a given price offer by the program in Fig. 4 with two segments.

The first segment starts at times 1, 2,. . . ,100, allowing 100 different potential contract cases to be investigated. We here determine the value of T as a value lying between 0 and 100. In order to illustrate the results, we also print the time and the value. In the second segment, we generate the bids that come at times 1.5, 2.5,. . . ,100.5, i.e., always a little later than when the value was determined. We leave

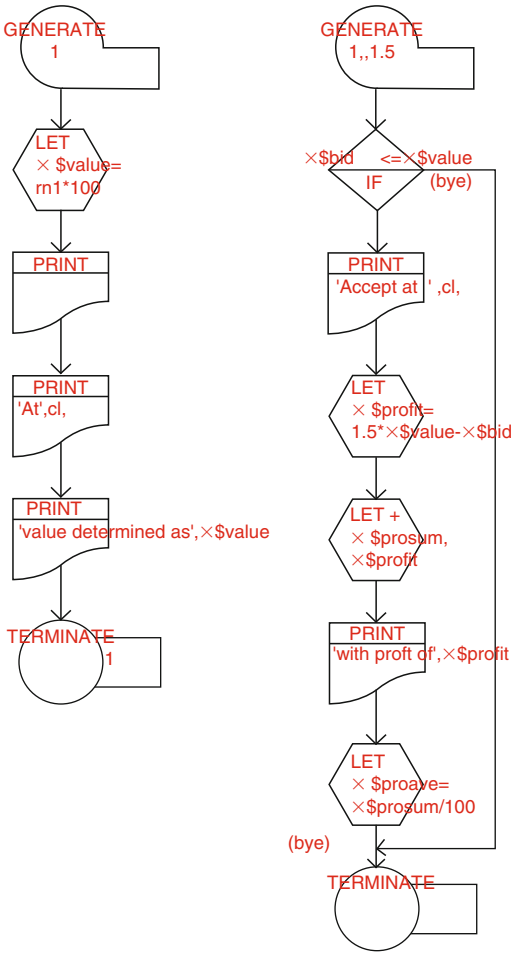


Fig. 4 Block diagram of bidding price model

with no contract if the bid is lower than the just determined value. Otherwise, we set the profit of A to 1.5 times the value *minus* the bid. We print that the value has been accepted and the resulting profit. We sum up all the profits and calculate the average of all profits. The program provides an output of the following sort for the case when 55 is A's bid. We include a few of the lines for illustration.

```
 At       1.00  Value  determined as     3.79
Accept at       1.50  With profit of  −49.32
 At      38.00  Value  determined as    14.96
Accept at      38.50  With profit of  −32.57
 At      39.00  Value  determined as    74.44
```

Already these few lines explain why there is a negative expected profit. In the cases when the value is larger than the bid, there is no acceptance. In the cases when

1.5*value is less than the bid there is a loss. There is a profit only in the cases when 1.5*value > bid > value. No matter the size of the bid, there is too large a probability that the value is outside the region defined by these inequalities.

To establish the optimal bid, we add the following line to the program:

```
help experi,x$bid,x$proave,10,0,90,20
```

This line implies that we run the simulation with 10 different values on the bid, ranging from 10 to 90, with 20 runs for each value, to get the results of the average profits. We then obtain the following output, showing that a 0 bid is optimal.

```
Invalue outvalue: limits with 95 % probab.
        Average    Lower limit   Upper limit
 0.00       0.00          0.00  —        0.00
 10.00     −0.21         −0.31  —       −0.12
                      . .
80.00     −16.96        −18.43  —      −15.50
90.00     −21.58        −23.39  —      −19.76
```

4 A Duopoly Game

We noted in the introduction that several simulation reports in the energy area dealt with market game situations. Our third example deals with a game based on DES, dealing with duopolies producing and selling homogenous goods, like oil or coal, but where demand is stochastic. Although the game is very small, it presents the main characteristics of games based on DES.

The basic feature of the game is that of a Bertrand duopoly game, where two firms sell identical products. If demand is deterministic, marginal costs are constant, production made to order, there are no limits to production capacities and there are no inventories, then the firm with the lower unit cost will in theory drive the price down to just below the unit cost of the competitor, who would then not be able to sell anything.

If one, however, assumes that demand is stochastic, that the marginal costs are not constant and that the firms produce to inventory, this conclusion does not hold. This was seen when playing the small DES game presented here in a set of experiments, run with students, where the prizes were financed by the research foundation of a major Swedish energy producer. The firm with the higher unit cost could survive by being able to sell at a higher price, since the other producer would run out of inventories from time to time, due to the stochastic demand variations, and the buyers would then purchase from the firm with inventories on hand, even if it charged a higher price.

There are also other factors making for a more complicated and realistic game. The firms do not make their decisions simultaneously, but are free to make them

at any time. Since each firm knows the decisions of the other firm, it is a game of perfect information. Furthermore liquidity is important, and payment behavior of customers is stochastic. There is an inventory carrying cost and a cost each time that a decision is made.

The game is run by two players, both sitting at the same PC. The computer first asks for the first decisions for the two players at time 0, i.e., at starting time. It prints some initial results which at this stage only show that both parties start with $200 in cash and that equity likewise is $200 for each firm, since they both start with no inventories. No profits have yet been made. Each player is now in turn asked to make three decisions: (1) the price to be quoted; (2) the number of units to be produced each month (there is a required production of one unit); and (3) the time until the next decision is to be made.

Assume that firm 1, which starts, sets the price to $10, the quantity to be produced to 10 units and the time until the next decision likewise to 10 days. This looks in the game dialog as follows:

```
For corp. 1: Give price (< $ 30)
10
Give monthly production (at least 1)
10
Give number of days until next decision
10
```

Let us assume that firm 2 sets all the three decision variables to 20. Since the next decision is to be made at time 10 by firm 1, the next thing the computer does is to give the following report at time 10 on firm 1's result.

```
Results for corp. 1 at time          10.00
 Profits          $            7.49
 Inventories      $             .00
 Bank             $          177.49
 Acc. receiv.     $           30.00
 Equity           $          207.49
```

The game then progresses with a succession of decision dialogs and result reports until either one firm goes bankrupt or a simulated year has passed.

In the GPSS program of the game, of around 150 blocks (Ståhl 1993), we first give the initial values for unit cost, costs of storing one unit for one month and the cost of making a decision. We define the functions for the annual sales quantities, each dependent on the price of the specific firm. The functions are of the constant price elasticity type with sales as a constant divided by price raised to the price elasticity constant. On the basis of these annual sales we calculate the *average* number of days between two orders for each firm. The actual time is obtained by multiplying this average by a sample from an exponential distribution with the average 1. These sales functions are used at the start of the sales segments, one for each firm.

Let us study that of firm 1. A GENERATE block generates one *potential* sales order after another for firm 1, the first coming at simulation start. This potential

order would constitute a true order for firm 1 only if firm 1 does not quote a higher price than firm 2. If firm 1 quotes a higher price, the order just generated is not valid, but thrown out. (Firm 1 might still make a sale, but in the corresponding firm 2 segment, if firm 2 is out of stocks.) If firm 1 has a lower price, the order is valid and we proceed to the following test. If firm 1 quotes exactly the same price as firm 2, firm 1 loses the order with 50% probability and wins it with 50% probability. If firm 1 wins the order, we test whether firm 1's stocks are empty. If this is the case, the buyer instead proceeds to purchase a unit from firm 2, if it has any in stock. However, if firm 2 is also out of stock, the buyer returns to firm 1 to wait for it to get a product into stock.

If firm 1 has no stocks, but firm 2 with the higher price has, the buyer will, however, buy from firm 2 only if he would be a buyer at this higher price that firm 2 is quoting. To handle this problem of a contingent demand curve, we use a stochastic technique. Assume that at present prices the annual demand facing firm 1 is 100, but for firm 2 only 80. Then 80% of the customers that arrive at firm 2 because firm 1 has run out of stocks would be willing to buy the product from firm 2. Thus for a customer arriving at firm 2, we sample the customer to stay and buy from firm 2 with 80% probability, while it with 20% probability goes back to firm 1 and waits for it to replenish its inventories.

The buyers who come back to firm 1 to wait until new products arrive from production segment 1, when they have either found firm 2 also to be without stocks or found the price of firm 2 too high, will only wait if they are certain to get deliveries within a month. We check that the number of customers waiting for deliveries is not more than the number of units to be produced in the next month. If this is not true, the newly arrived customer leaves the system; otherwise he waits until new products arrive.

At the actual sales part for firm 1, the buyer takes one unit out of stock. Next cumulative profits increase by the price of the sold unit and decrease by the unit cost of the sold unit. Next, simulation time is advanced by a sampled time to reflect the credit time used by the buyer. Average credit time is 30 days, but some buyers pay within a shorter time and some pay within a considerably longer time. We here sample from the Erlang distribution with a shape factor of 3. After this, payment is received and cash increases by the product price.

The *production segment* is very simple. Firm 1 produces at a monthly rate of PROQ1, i.e., delivers a product every 30/PROQ1 days, without any stochastic variations. The production of a unit implies that we increase our inventory with one unit and cash is decreased by the unit cost of production. Payments are made directly. If cash (checking account balance) then becomes negative, the firm tries to borrow. If it needs to borrow more than the maximum borrowing limit, the corporation goes bankrupt. At bankruptcy, we get a report on the corporation going bankrupt and the simulation is stopped.

The report and decision segment for firm 1 starts with a block that initiates the first report and decisions at time 0 and then is repeated each time firm 1 is to make a new decision. The cost of holding inventories is the present inventory level times the monthly inventory cost per unit multiplied by the time since the last decision,

measured in months. We next add the interest costs (or deduct interest income). The interest is calculated for the period since the last decision. All costs are deducted from total equity as well as from cash. Profits are then calculated as present equity minus equity at the time of the preceding report. In the report, we also print the value of stocks, the amount of cash (if negative = overdraft) and the value of accounts receivable. Finally, equity is printed.

Before this segment asks for new decisions, the program checks that equity is not negative and that the overdraft does not exceed the credit limit. If this is not true, bankruptcy is declared. Otherwise the value on price is input. The price is restricted, since a high price might cause too long a time between the orders. Next firm 1 inputs the monthly production rate, and finally the time until the firm 1 is going to make the next decision. The program then schedules the next report and decision for firm 1 to take place at present time + the just input time delay. At that time, the program goes through another round of reports and decisions for firm 1.

Similar sales, production and report segments apply to firm 2. On day 360 the stop segment finally stops the game, with a final report for both players, if bankruptcy has not occurred earlier.

This game can easily be extended in several ways. One can vary the game parameters, like costs, price elasticity, initial cash, credit limit, etc. One can also include more than two players or a more complicated production segment with a more elaborate cost function. A major change would be to include more decision variables influencing sales, like advertising. The demand function would then be more complicated, and it is questionable that a DES game would be suitable.

A simpler and probably a more interesting change would be the introduction of a *robot* player, i.e., to make the game into one, playable by only one person playing the role of firm 1 against a very simple 'robot' handling firm 2. The only differences compared to the game above refer to the decisions in the report and decision segment. With a constant unit cost, c, we could, using traditional optimization, have an optimum when $MR = MC = c$. In this case of a constant (absolute value of) price elasticity e, the robot would set the price of firm 2, p_2, as $c(1 + 1/(e\text{-}1))$. This is the optimal price provided that this is the lower price and stochastic variations are disregarded. However, if this optimal price $p_2 > p_1$, then firm 2 will instead undercut firm 1, e.g., set $p_2 = p_1 - 0.1$. The monthly production is then set by the robot as the annual sales at this price divided by the 12 months of the year. The time of the next decision for firm 2 is set on basis of the next time of decision of firm 1. The idea is that firm 2 should try to always make its decisions immediately after firm 1. In this way, firm 2 can always immediately undercut the price that firm 1 has just set. This would be an optimal strategy for firm 2, provided firm 1 does not make a lot of decisions.

Summing up, the simple DES duopoly game will, even in its simplest form, give some insights into how stochastic variation in key variables can lead to results that are at odds with the results of the traditional deterministic Bertrand model. The DES game can also easily be extended to allow for further use for experiments and in education.

References

Bompard, E., & Abrate, G. (2007). *Game theory models for electricity market simulations*. Working paper, Politecnico de Torino.

Born, R., & Ståhl, I. (2007). *Simulation made simple with WebGPSS*. Gothenburg: Beliber.

Born, R., & Ståhl, I. (2008). A business course in simulation modeling. *Issues in Information Systems, IX*(1), 6–15.

Brady, T. (2001). Computer simulation analysis of electricity rationing effects on steel mill rolling operations. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 Winter Simulation Conference* (Vol. 1, pp. 946–948). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc01papers/125.PDF. Accessed 26 January 2009.

Chandrakanthi, M., Hettiaratchi, P., Prado, B., & Ruwanpura, J. Y. (2002). Optimization of the waste management for construction projects using simulation. In E. Yücesan, C. H. Chen, J. L. Snowdon, & J. M. Charnes (Eds.), *Proceedings of the 2002 Winter Simulation Conference* (Vol. 2, pp. 1771–1777). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc02papers/077.pdf. Accessed 26 January 2009.

Conner, S. L., Lee, J., & Okoye, C. (1993). A simulation model for analysis of long term natural gas commitment. In *Proceeding of the 1993 Winter Simulation Conference* (pp. 1372–1373).

Halpin, D. W., & Martinez, L. H. (1999). Real world applications of construction process simulation. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 Winter Simulation Conference* (pp. 956–962). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc99papers/138.PDF. Accessed 26 January 2009.

Hamblin, D. M., & Ratchford, B. T. (2002). Batting average: a composite measure of risk for assessing product differentiation in a simulation model. In E. Yücesan, C. H. Chen, J. L. Snowdon, & J. M. Charnes (Eds.), *Proceedings of the 2002 Winter Simulation Conference* (Vol. 2, pp. 1578–1587). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc02papers/216.pdf. Accessed 26 January 2009.

Ioannou, P. G. (1999). Construction of a dam embankment with nonstationary queues. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 Winter Simulation Conference* (pp. 921–928). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc99papers/133.PDF. Accessed 26 January 2009.

Jacinto, C. M. C. (2002). Discrete event simulation for the risk of development of an oil field. In E. Yücesan, C. H. Chen, J. L. Snowdon, & J. M. Charnes (Eds.), *Proceedings of the 2002 Winter Simulation Conference* (Vol. 2, pp. 1588–1592). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc02papers/217.pdf. Accessed 26 January 2009.

Laughery, R., Plott, B. M., Engh, T. H., & Scott-Nash, S. (1996). Discrete event simulation as a tool to determine necessary nuclear power plant operating crew size. In *Proceedings of the 1996 Winter Simulation Conference* (pp. 1272–1279).

Kaproń, H., & Kaproń, T. (2007). *Modelling of energy market processes in the distributor monopolist structure*. Working paper, Lublin University of Technology.

Kauffmann, P. J. (1998). Using simulation with a logit choice model to assess the commercial feasibility of an advanced environmental technology. In D. J. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 Winter Simulation Conference* (Vol. 2, pp. 1513–1518). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc98papers/206.PDF. Accessed 26 January 2009.

Mallor, F., Azcarate, C., & Blanco, R. (2007). Economic assessment of energy systems with stimulation and linear programming. In *2007 Winter Simulation Conference*. doi: 10.1109/WSC.2007.4419877.

Nolan, P. J., O'Kelly, M. E. J., & Fahy, C. (1993). Assessment of ways of improving the supply continuity in electric power systems – a simulation approach. In *Proceedings of the 1993 Winter Simulation Conference* (pp. 1192–1200).

Raychaudhuri, S. (2007). *Development of a web-based simulation game of oil supply chain*. Paper presented at 2007 Informs Conference.

Samuelson, W. F., & Bazerman, M. H. (1985). Negotiation under the winner's curse. In V. Smith (Ed.), *Research in experimental economics* (Vol. 3, pp. 105–137). Greenwich, CT: Jai.

Sanders, R. L., & Lake, J. E. (2005). Training first responders to nuclear facilities using 3-D visualization technology. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, & J. A. Joines (Eds.), *Proceedings of the 2005 Winter Simulation Conference* (pp. 914–918). Resource document. INFORMS Simulation Society. http://www.informs-sim.org/wsc05papers/107.pdf. Accessed 26 January 2009.

Shi, J., & AbouRizk, S. M. (1994). A resource based simulation approach with application in earthmoving/strip mining. In *Proceedings of the 1994 Winter Simulation Conference* (pp. 1124–1129). doi: 10.1109/WSC.1994.717498.

Ståhl, I. (1993). A small duopoly game based on discrete-event simulation. In A. Pave (Ed.), *Modelling and simulation ESM. 1993*. Lyon: SCS.

Sturgul, J., Lorenz, P., & Osterburg, S. (2001). Simulation in the mining industry. In T. Schulze, S. Schlechtweg, & V. Hinz (Eds.), *Simulation und Visualisierung 2001* (pp. 1–16). Magdeburg, März 2001.