International Institute for
Applied Systems Analysis
IIASA www.iiasa.ac.at

Natural Environment Management and Applied Systems Analysis

Makowski, M. and Nakayama, H.

IIASA Interim Report
July 2001

International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report **IR-01-021**

Natural Environment Management and Applied Systems Analysis

Marek Makowski (marek@iiasa.ac.at)
Hirotaka Nakayama (nakayama@konan-u.ac.jp)

Approved by

Arne Jernelöv
Acting Director, IIASA,

July 2001

Natural Environment Management and Applied Systems Analysis

Proceedings of Konan–IIASA Joint Workshop, September 6–8, 2000

Marek Makowski and Hirotaka Nakayama
Editors

IR-01-021 / July 2001

International Institute for Applied Systems Analysis
A-2361 Laxenburg, Austria

Hirao Taro Foundation of the Konan University Association
for Academic Research, Kobe, Japan

The International Institute for Applied Systems Analysis

is an interdisciplinary, nongovernmental research institution founded in 1972 by leading scientific organizations in 12 countries. Situated near Vienna, in the center of Europe, IIASA has been producing valuable scientific research on economic, technological, and environmental issues for nearly three decades.

IIASA was one of the first international institutes to systematically study global issues of environment, technology, and development. IIASA's Governing Council states that the Institute's goal is: *to conduct international and interdisciplinary scientific studies to provide timely and relevant information and options, addressing critical issues of global environmental, economic, and social change, for the benefit of the public, the scientific community, and national and international institutions*. Research is organized around three central themes:

- Energy and Technology;
- Environment and Natural Resources;
- Population and Society.

The Institute now has National Member Organizations in the following countries:

Austria
The Austrian Academy of Sciences

Bulgaria*
Ministry of Environment and Waters

Czech Republic
The Academy of Sciences of the Czech Republic

Finland
The Finnish Committee for IIASA

Germany**
The Association for the Advancement of IIASA

Hungary
The Hungarian Committee for Applied Systems Analysis

Japan
The Japan Committee for IIASA

Republic of Kazakhstan*
The Ministry of Science – The Academy of Sciences

Netherlands
The Netherlands Organization for Scientific Research (NWO)

Norway
The Research Council of Norway

Poland
The Polish Academy of Sciences

Russian Federation
The Russian Academy of Sciences

Slovak Republic*
The Executive Slovak National Committee for IIASA

Sweden
The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS)

Ukraine
The Ukrainian Academy of Sciences

United States of America
The American Academy of Arts and Sciences

*Associate member
**Affiliate

Foreword

This volume is composed of revised papers that were accepted for the NEMASA, Konan-IIASA Joint Workshop on Natural Environment Management and Applied Systems Analysis that was held on September 6–8, 2000, at IIASA, Laxenburg, Austria. The workshop is part of the activities of the research project *Modeling by Computational Intelligence and its Application to Natural Environment Management*, which is being supported by the Hirao Taro Foundation of the Konan University Association for Academic Research, Kobe, Japan.

The management of the natural environment, especially in practicing advanced agriculture, is one of the challenging problems faced by modern societies. Many of the techniques in applied systems analysis hold promise for working out this problem. The purpose of this workshop was to present new concepts and methodologies for managing the environment, and to offer an open forum for exchanging ideas between various research disciplines; especially, between agro-environmental and applied systems analysis research, and between researchers and practitioners.

The papers deal with a range of topics. We have arranged them into the following categories: (1) modeling methodologies, (2) data analysis, (3) land use, (4) water management, and (5) applications. The paragraphs that follow discuss the placement of each chapter in this overall scheme.

1. *Modeling Methodologies*. The chapters in this part present various modeling paradigms that are illustrated by using real-world applications. In Chapter 1, J. Wessels provides an overview of the types of models used in natural resources management by formulating a rough categorization of decision problems and providing many examples. In Chapter 2, J. Sendzimir describes a process, called Adaptive Environmental Assessment, that has developed over 30 years of experiments about abilities to integrate inquiry, understanding, and actions in the face of surprising shifts in evolving natural resource systems. M. Makowski, in Chapter 3, presents selected modeling paradigms applied to model-based decision support; these paradigms are illustrated by

discussing their applications to the RAINS model, which is a complex model for analysis of cost-effective policies aimed at improving European air quality. The other three chapters in this part deal with applications of two novel methodologies. Chapter 4, by K. Hayashi, reviews two methodologies used for analysis of agro-environmental problems, especially for evaluating agricultural practices, namely multicriteria analysis for selecting farming practices, and risk analysis for health and ecological issues. In Chapter 5, P. Heiskanen presents a constraint proposal method applied to international negotiations aimed at improving air quality, and using the simplified RAINS model. S. Stagl and coworkers, in Chapter 6, show how a particular multicriteria decision aid method, called NAIADE, has been applied to ranking alternative projects in a large-scale ecosystem protection program.

2. *Data Analysis*. The chapters in this part involve methodological issues and applications of data analysis. Z. Pawlak's Chapter 7 introduces Rough Sets, which is a novel but already well-established approach to data analysis, and uses simple examples to illustrate its applicability to complex problems. The next two chapters deal with DEA (Data Envelopment Analysis). Y. Yun and coworkers, in Chapter 8, present a generalization of several DEA methods applied to multi-criteria decision analysis. In Chapter 9, P. Korhonen and M. Luptacik show, working with the example of analysis of eco-efficiency of power plants, how DEA can be used to facilitate public discussion on environmental policies. Chapters 10 and 11 discuss novel methods and techniques of data analysis. First, E. Watanabe and coworkers show how a specialized technique involving neural networks was applied to the time-series prediction for analyzing an hourly traffic problem. Second, M. Tanaka and M. Asada present a non-linear regression analysis problem for which neural networks are not suitable, and to which another regression analysis technique has been successfully applied.
3. *Land Use*. This part is composed of four chapters that deal with land-use problems. In Chapter 12, K. Hubacek and L. Sun present the problems of land demand and supply in China, which are of crucial importance for China's development owing to the fast economic growth, urbanization, changes in life style, and population growth. Next, in Chapter 13, A. Mohamed describes an approach for integrating agro-ecological and agro-economic analysis related to land-use planning and its connection with land-use policy options. In Chapter 14, Y. Takahashi discusses the importance of cattle grazing in land resource management, illustrating his points with the case study of the Mt. Sanbe area. In Chapter 15, M. Tiongco shows how the soil-quality index can be estimated so that it accounts for technical efficiency of agricultural land use in the Philippines.

4. *Water Management*. The four chapters of this part deal with water management problems. First, in Chapter 16, G. Fischer and D. Wiberg present a large case study on the impacts of climate change on water resources in China and their relations to economic and environmental factors. In Chapter 17, M. Grauer and coworkers apply a novel approach to ground water management and the implementation of advanced computational techniques to a case study in Germany. J. Pakulska presents, in Chapter 18, the changes in the water management system in Poland that are necessary for dealing with problems of sustainable environment management after the transition of the Polish economy. In Chapter 19, P. Bartoszczuk describes a model for price setting proposed to be applied to pricing of municipal water supply in Poland.
5. *Applications*. In the final part, four chapters present diversified applications and approaches to various problems of natural environment management. In Chapter 20, W. Ciechanowicz presents plans developing bioenergy in the transport sector and shows how this could contribute to solving some of the rural development problems in Poland. R. Cumpston describes novel techniques for the development of regional population projections in Australia in Chapter 21. The last two chapters deal with forest-related problems. M. Flinkman and coworkers, in Chapter 22, present the methodology for identifying practices essential to the development of sustainable forest management in the Siberian forest, using the rough-set analysis. In Chapter 23, M. Obersteiner demonstrates an innovative approach based on auction theory to analyze Siberian forest sector during its transition from a centrally planned economy to an economy guided by market principles.

Marek Makowski
Hirotaka Nakayama

Acknowledgments

The editors of this volume were also the organizers of the NEMASA workshop and wish to thank IIASA for providing facilities and organizational help for this workshop. We gratefully acknowledge financial support by the Hirao Taro Foundation of the Konan University Association for Academic Research not only for supporting the NEMASA workshop, but also for publishing these proceedings.

We would like to thank all the authors for their cooperation in the requested modifications of the papers, and to Ms. E. Delpos, Ms. M. Elliott, Ms. A. James, and Ms. C. Kugi of IIASA's Publications Department for their hard work to convert the diversified manuscripts into the nice form presented in this volume. Without their cooperation it would not have been possible to prepare this volume in a relatively short period of time.

Contents

Part I: Modeling Methodologies

Chapter 1

Decision Support for Natural Resource Management Models and Evaluation Methods

Jaap Wessels

Abstract

When managing natural resources or agrobusinesses, one always has to deal with autonomous processes. These autonomous processes play a core role in designing model-based decision support systems. This chapter tries to give insight into the question of which types of models might be used in which cases. It does so by formulating a rough categorization of decision problems and providing many examples. Particular attention is given to the role of statistical learning theory, which may be used to replace mathematical modeling by training with examples.

Keywords: Decision support systems, natural resource management, mathematical programming, agromanagement, statistical learning theory.

1.1 Introduction

When speaking about decision support and decision analysis, it is important to restrict the subject, since decision support problems may have very different natures. For some decision problems it is most essential to structure the decision making process by indicating who should decide on which aspect and when. In other decision problems, it is most urgent to provide well-structured information about the current situation and possibly about the past. However, in natural resource management and in agromanagement, high priority is assigned to forecasting the consequences of possible decisions. Therefore, in this type of decision making, modeling is an essential feature, since models may be used to provide information about the consequences of possible decisions. This focus does not imply that structuring of the decision process and providing well-structured information are irrelevant. However, this chapter will concentrate on the modeling aspect.

When managing natural resources or agrobusinesses, the underlying processes are always relatively complex and, therefore, one needs models in order to obtain insight into the relationships between decisions and consequences. The only alternative might be to rely on methods from artificial intelligence or statistical learning theory. Such methods exploit experience in previous related cases or the knowledge of experts. In Section 1.9, I will return to this possibility, particularly to the use of statistical learning theory. In the other sections, I will primarily deal with modeling approaches.

Section 1.2 explains how the relevant decision problems may be categorized. The subsequent six sections each treat one category of decision problems. Each category is roughly outlined and mainly clarified by examples.

1.2 The Modeling of Decision Problems

Decision problems about natural resources or agrobusinesses are always related to underlying processes that are highly autonomous. Such processes can be of a physical, chemical, biological, demographic, economic, or technical nature. A typical example is the spreading and transformation of air pollutants by wind and sun. Another example is the growing of wheat under the influence of soil and weather. With respect to the air pollution example, decisions can only affect the emissions, but, once they are emitted, one must take the processes leading to deposition for granted. With respect to the wheat growing example, the farmer may affect the starting conditions by selecting the right seed and preparing the soil in the proper way, but afterwards his influence is restricted.

When making decisions, one has to take these largely autonomous processes into account, since the consequences of decisions are generated through these processes. Therefore, when discussing the modeling, these autonomous processes play a central role. It even seems natural to take the role and type of these processes as the basis for categorizing decision problems.

The reason for modeling is to obtain insight into the relations between possible decisions and consequences. Therefore, these relationships determine which processes should be modeled with which level of detail. Naturally, one also has to regard the possibilities of evaluating models. Hence, a compromise might be necessary.

If one considers air pollution, then one is interested in a chain of processes. The first link involves considering the processes that generate the emissions, like driving cars and producing electricity; the second link constitutes the technical, physical, and chemical processes of emission, transport, transformation, and deposition of pollutants; the final link involves the processes that represent the impact of pollutants on human health, quality of trees, etc. For making *decisions* on emissions, however, one may argue that a description of consequences in terms of depositions and air quality is sufficient. Such a conclusion obviates a lot of tedious modeling: one only needs a model that translates economic, technical, and demographic activities in emissions and a model that translates emissions in depositions and air quality characteristics. Thus we arrive at a kind of modeling in which the natural resources don't explicitly appear. And this is a quite common procedure if one considers large-scale environmental decision problems. This situation describes our first category of decision problems, as set forth with more examples in Section 1.3. In the subsequent sections, living creatures play an increasingly explicit role.

In Section 1.4, I consider problems where the behavior of living creatures is essential and in the sections that follow, the life cycles of animals or plants form the starting point for modeling. In Section 1.5, I analyze decision problems where life cycles generate tasks that have to be performed effectively and efficiently. In Section 1.6, I consider decision problems regarding the starting or stopping of life cycles. Section 1.7 involves decision problems that concern starting or side conditions that affect the proceeding of life cycles. Finally, Section 1.8 treats problems in which life cycles may be influenced dynamically.

1.3 Decision Problems Without a Direct Relationship to Living Creatures

Wierzbicki *et al*. (2000) give an extensive treatment of decision support for environmental problems. The cases treated there all belong to the category described in this section. As previously explained for the case of air pollution, there are good arguments for separating studies on the impact of air pollution from studies on depositions and air quality. For the latter type of studies, we are typically dealing with "physical" laws regarding emissions, transportation, transformation, and deposition of pollutants. Here, demographic and economic processes are described in the same way as truly physical processes.

For an extensive treatment of the modeling of such problems, the reader is referred to Wierzbicki *et al*. (2000). Here I simply give some examples to clarify what type of problems fall into this category and what types of models are relevant. A common feature of these examples is that all regard policy making on a higher political level.

Examples:

a. *Transboundary air pollution.* In Europe, air pollution is an international problem, since some countries suffer more from emissions by other countries than from their own emissions. The RAINS-model of IIASA has been developed to support negotiations between European countries regarding abatement measures. The RAINS-model is one of the rare examples of a mathematical model being accepted as the basis for negotiations.

   The RAINS-model is a mathematical programming model with a large linear part, but also with a substantial nonlinear part caused by the generation process of tropospheric ozone. For algorithmic reasons, the model contains considerable simplifications like yearly averages and simplified sources.

   For a more extensive treatment and several references, see Amann and Makowski (2000), and Chapters 3 and 5 of this book.

b. *Energy planning*. There are many decision problems regarding generation and distribution of energy. For environmental reasons, medium-term and long-term decisions are particularly relevant. Several international bodies are involved in studies and negotiations between countries regarding energy supply and utilization. In such studies, linear programming models play an important role. These models provide a rather direct translation from reality. For an overview, see Messner *et al*. (2000).

c. *River basin water quality*. In river basins, the water is polluted by some players and used by others. It even occurs that several players pollute the water and are in extreme need of clean water at the same time.

   Several measures may be taken to improve the overall water quality. However, such measures are usually expensive and may have unpleasant side effects on the economy. Makowski and Somlyody (2000) show how such a decision problem may be supported by a mixed-integer linear programming model. This model uses a simplified version of the detailed model describing the transportation and transformation of pollutants in a river basin.

d. *Land use planning*. Different ways of using land compete for this scarce resource. Moreover, the way land is used has a considerable impact on the food supply, the water availability, and on several other important issues. Fischer and Makowski (2000) describe how linear programming models may support an integrated approach towards land use planning. In this volume, Fischer and Wiberg consider the possible impacts of climate change on water-stressed agriculture in Northeast China (see Chapter 16).

e. *Groundwater management*. Changes in groundwater level may have a considerable impact. Therefore, it is necessary to perform relatively detailed studies on groundwater in case of infrastructural operations which might affect the groundwater level in the neighboring area. Grauer *et al*. (Chapter 17 in this volume) provide a solution by coupling an optimizing algorithm to a simulation model based on finite elements. The computational complexity is beaten by using distributed computations.

A major problem in all these examples is their size, which, in some cases, is substantially diminished by simplifying process models considerably.

1.4 Behavioral Models

If living creatures are involved in the decision problem, then, usually, their life cycles provide the basic information for modeling. However, in rare cases, the primary source of modeling information is the behavior of animals. We give one example of such a case.

Example:

a. *Design of robotic dairy barns*. The most up-to-date dairy barn is equipped with one or more milking robots. The main advantage of milking robots over conventional milking machines is that cows may go for milking more than two times a day, which gives a considerable increase in milk yield. A dairy barn consists of different resources and the design problem is to find a good balance between numbers and sizes of the different resources. The needs are determined by the frequencies of visits and the time spent per visit.

   Halachmi *et al*. (2000) present a decision support system based on a queueing network model for the behavior of the cows.

1.5 Life-Cycle Generated Tasks

In several operational planning problems in agriculture, the life cycles are no longer influenced, but they do generate tasks which have to be performed effectively and efficiently. The nature of the products quite often dictates that tasks be executed quickly after they are generated.

Examples:

a. *Internal transport in pot plant nurseries*. Modern pot plant nurseries have specialized working areas for activities like potting, sorting, spacing, harvesting, and growing, since they apply dedicated equipment for each of these activities. Therefore, a lot of internal transport is necessary, which requires decisions regarding lay-out, transport equipment, allocation, and sequencing. Annevelink (1999) deals with the operational aspects of transportation in pot plant nurseries. He recommends a combination of simple rules for parking with the use of local search techniques like simulated annealing, tabu search, and genetic algorithms for the sequencing.

b. *Scheduling of inseminations*. Inseminators travel to the farms where cows are to be inseminated with the sperm of a bull selected by the farmer. The farmer calls for an insemination when s/he thinks that it is the right time for a particular cow. S/He also asks for sperm of a particular bull from the catalogue. For various reasons there is a tendency to use fresh rather than frozen sperm. Two times a day, farms should be assigned to inseminators and a route should be determined for each inseminator. Different techniques are in use for these purposes.

   Also the inventory management of sperm provides interesting decision problems. The amount produced cannot be affected on short notice, but it should be decided for each bull which fraction should be frozen and how much fresh sperm should be dispatched to the regional subdepots.

c. *Dealing with manure*. Due to legal restrictions, manure may only be used in a restricted way in The Netherlands. These legal restrictions are based on conventions of the European Union. Because of the wide-spread bio-industrial activities in The Netherlands, particularly pig-breeding and poultry-keeping, these restrictions have much more impact than in most other EU-countries. Non-used manure should be processed or transported to other areas for controlled application. Processing and transportation are expensive for the farmers and direct application is only allowed to a restricted level. There are several decision problems related to dealing with manure.

   For strategic and tactical decisions on a regional scale, a decision support system has been developed (compare De Mol and Van Beek, 1991). This system primarily uses linear programming. For some extensions, mixed-integer linear programming is used.

d. *Logistics of biomass collection*. Biomass may be used as fuel in energy plants. One of the main cost factors for biomass energy production is the cost of transportation and handling. Biomass for energy production may stem from several sources, e.g., restproducts (like demolition wood and waste paper), agricultural by-products (like straw and tops) and crops which are specifically cultivated for energy production (like willow and poplar). De Mol *et al*. (1997) show that mathematical models can help in designing an efficient logistic structure for collecting biomass. The authors present some models of their own and review the literature on the topic. Their paper shows that several types of models may be useful for different types of decisions. These models range from simulation models through dynamic programming models to mixed-integer linear programming models.

e. *Design and management of distribution centers for perishables*. Perishables, like fruits and vegetables, generate special questions regarding the design and management of distribution centers with respect to stock allocation, inventory policies, lay-out etc.. Broekmeulen (1998) shows that local search methods may be used profitably for assignment of perishables to zones, for stock allocation, and for some other operational decision problems. For some other decision problems, stochastic dynamic programming and linear programming appear to be useful.

One conclusion we may draw from this set of examples is, namely, that explicit modeling of life cycles is nearly never needed in this type of decision problem. In the subsequent sections we will consider problems in which life cycles play an increasingly explicit role.

1.6 Decisions Regarding Starting and/or Stopping of Life Cycles

It is quite common that the proceeding of life cycles is only affected by the decision when they should start and when they should stop. Determining seeding and harvesting times are major decisions in agriculture. But determining which type of product should be seeded is also an important decision.

Examples:

a. *Crop selection*. There are different reasons why crop selection at the level of individual farms may be a complex problem. A first reason may be restrictions on the order of particular crops in order to avoid plant diseases and soil quality deterioration. A second reason may be the restricted availability of resources. A third reason may involve time restrictions with respect to the seasons. A fourth reason may be the risks with respect to prices, weather, and plant diseases.

   Models that are used for these types of problems are linear programming, mixed-integer linear programming, and stochastic programming models.

b. *Timing of insemination through estrus detection*. As explained in Example b of Section 1.5, the dairy farmer must determine when a cow is ready for insemination. For the milk yield it is important that the insemination has a high probability of success and that no opportunities are overlooked. The most important determinant of the success probability is the timing of the insemination. Usually, the farmers determine the right time by observing the cow. De Mol (2000) developed a method for automatic detection of the right time for insemination (estrus) of dairy cows. In a modern dairy barn (compare Example a of Section 1.4), the behavior of the cows can be observed continuously. For instance, the milk yield and milk temperature are measured, but also the intake of concentrated food and the tendency to roam. Using the time series of such measurements and a few others, De Mol applies a Kalman filter approach for forecasting the time of estrus.

c. *Determining harvesting strategies for fisheries*. In natural environments it is important to keep sufficient fish stock for procreation and for prey (e.g., for other types of fish or for birds). To determine good harvesting strategies (locations, timing, and quantities), a model of the life cycle is necessary. Such a model should at least include the interaction between growth, procreation, food availability, and other environmental aspects. Models exist for different types of fish and shellfish (see, e.g., Scholten and Smaal, 1999, for such a model for mussels). These models may be used for supporting scenario analyses.

   In fishing nurseries, it is particularly important to find a good balance between food, growth, and prices. Here linear programming is used, but also (stochastic) dynamic programming.

In these examples, we see an increasing need to use life cycle models. In the next section, a type of problem will be presented that requires more detailed models of (parts of) life cycles.

1.7 Decisions About Start and/or Side Conditions for Life Cycles

Problems become more complicated if one tries to influence start and/or side conditions for life cycles in order to affect their proceedings. One simple example involves decisions regarding the preparatory work before seeding. Two other examples follow.

Examples:

a. *The choice of the right bull-cow combination*. Farmers consult the performance indicators of the available bulls in the catalog at the insemination station when choosing sperm for their cows. In practice, very few bulls appear to be favorite sperm providers for Frisian-Holstein cattle worldwide. In fact, all Frisian-Holstein bulls and cows belong to one genetic line. For instance, the popular bull Sunny Boy has about a million offspring. This situation poses a considerable risk of increase in inbreeding. Bijma *et al*. (2000) provide a general procedure for predicting rates of inbreeding. This procedure can be used to decide to avoid the sperm of certain bulls for a particular cow.

b. *Improvement of a population*. Apart from possible harvesting and predation losses, a population of fish, shellfish, mammals, or birds is affected by climatic circumstances (e.g., water temperature), physical environment (e.g., water flows) and food availability. These circumstances may be affected to some extent – deliberately as well as by happenstance. The consequences of changes may be evaluated by using a life cycle model which includes the relation between growth, food availability, and the reproduction success rates as a function of the circumstances. For an example of such a model, see Scholten and Smaal (1999).

As would be expected, these examples require rather detailed models of (aspects of) life cycles.

1.8 Problems in Which Life Cycles May Be Affected Dynamically

Environmental or agricultural management often reacts to the state of life cycles. However, decision support for problems of that operational management type is rare.

Examples:

a. *Operating a cut flower nursery under dynamic demand and price.* In cut flower nurseries, the growth can be speeded up or retarded to some extent. For instance, some flowers need a cold period before they are willing to blossom. By putting them in a freezer for some time and in a hothouse afterwards, the time of blossoming may be influenced. Good timing may have a considerable influence on the price, but it also affects costs.

   For a decision support system for this purpose, one needs a model of the relationship between growth and temperature profile and also a dynamic forecasting procedure for market prices, since prices of flowers are affected by the weather and by some other dynamic features.

b. *Operational management of commercial woods*. The growth of trees is largely determined by dynamic features like weather, diseases, and tree density. Operations like thinning and harvesting can be based on the actual situation as measured by remote sensing or aerial photographs. Also availability of resources is a relevant constraint. Different types of models are used, ranging from linear programming to (stochastic) dynamic programming.

Here we conclude the overview of models based on the way life cycles play a role in the modeling.

1.9 Statistical Learning

In the preceding sections, the emphasis was on explicit modeling of relationships that were supposed to be important for making decisions. However, explicit modeling is not always possible, particularly where relationships are complex and not well-understood. In such cases *statistical learning techniques* may replace explicit modeling. Statistical learning techniques make a systematic use of experience in related cases. In practice it has appeared that statistical learning techniques may be useful for recognizing patterns. This facility may be applied to performing classification tasks and also to estimating response functions. Clearly, this approach only works if enough experience in related cases is available.

Examples of statistical learning techniques include:

i. *Neural nets*. Among several variants of neural nets, we mention:

   - multi-layered perceptrons,
   - Hopfield networks, and
   - self-organizing maps.

ii. *Support vector machines*. For further information on this topic, the reader is referred to Vapnik (2000).

When managing natural resources or agricultural systems, three possible roles exist for statistical learning techniques. Below, we explain each of these roles and provide an example of each:

1. *Interpretation of observations or measurements*. Many situations present a lot of data that require interpretation. If many data points exist, which are already associated with an interpretation, then it may be attractive to train a neural net or other statistical learning technique as an interpreter.

   Example:
   Translation of remote sensing data of woods in operational characteristics may be used in Example b of Section 1.8. For an example of classifying remote sensing data with different types of neural nets, see Suurmond and Bergkvist (1996).

2. *Forecasting of time series*. There exist several decision problems for which forecasting of time series is an essential part. Particularly in cases where modeling seems to be difficult, statistical learning techniques provide an alternative.

   Example:
   For detection of estrus or mastitis of dairy cattle, forecasting of time series is essential (compare Example b of Section 1.6).

   By using Kalman filters a relatively rigid model is chosen [see De Mol (2000)]. Statistical learning might provide a more flexible class of relationships.

3. *Suggesting decisions*. If it afterwards becomes clear which decision should have been taken, it is possible to collect a set of learning pairs, consisting of a possible situation and the corresponding desirable decision. Particularly if it is difficult to provide a model which generates the decisions, it is attractive to use the learning pairs for the training of some statistical learning technique.

   Example:
   When determining market strategies in a market with a high price variability (like nursery products, fish, potatoes), it may be attractive to avoid explicit modeling and train some statistical learning technique instead.

The future will show what kind of position statistical learning techniques will obtain in management of natural resources and agricultural businesses.

1.10 Final Remarks

It would have been possible to introduce another approach to partitioning decision problems for managing natural resources and agricultural systems. However, every partitioning has its weak sides. With the presented partitioning and the large number of examples, I hope to have shown how rich the set of relevant decision problems is and how effective model-based decision support can be for taking well-founded management decisions in this area.

Quite a few of the references noted here are related to research with which the author has some familiarity, executed either at IIASA in Laxenburg, Austria, or in the Netherlands. Most of the publications cited contain ample references to related work elsewhere.

References

Amann, M., and Makowski, M., 2000, Effect-focused air quality management, in A.P. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 367–398.

Annevelink, E., 1999, Internal Transport Control in Pot Plant Nurseries, Ph.D. thesis, Wageningen Agricultural University, Wageningen, Netherlands.

Bijma, P., Van Arendonk, J.A.M., and Woolliams, J.A., 2000, A general procedure for predicting rates of inbreeding in populations undergoing mass selection, *Genetics*, **154**:1865–1877.

Broekmeulen, R.A.C.M., 1998, Layout and Operations Management of Distribution Centers for Perishables, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, Netherlands.

De Mol, R.M., 2000, Automated Detection of Oestrus and Mastitis in Dairy Cows, Ph.D. thesis, Wageningen Agricultural University, Wageningen, Netherlands.

De Mol, R.M., and Van Beek, P., 1991, An OR contribution to the solution of the environmental problems in the Netherlands caused by manure, *European Journal of Operational Research*, **52**:16–27.

De Mol, R.M., Jogems, M.A.H., Van Beek, P., and Gigler, J.K., 1997, Simulation and optimization of the logistics of biomass fuel collection, *Netherlands Journal of Agricultural Science*, **45**:219–228.

Fischer, G., and Makowski, M., 2000, Land use planning, in A.P. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 333–365.

Halachmi, I., Adan, I.J.B.F., Van der Wal, J., Heesterbeek, J.A.P., and Van Beek, P., 2000, The design of robotic dairy barns using closed queueing networks, *European Journal of Operational Research*, **124**:437–446.

Makowski, M., and Somlyody, L., 2000, River basin water quality management, in A.P. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 311–332.

Messner, S., Strubegger, M., and Wierzbicki, A.P., 2000, Energy planning, in A.P. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 399–414.

Scholten, H., and Smaal, A.C., 1999, The ecophysiological response of mussels in mecocosms with reduced inorganic nutrient loads: simulations with the model EMMY, *Aquatic Ecology*, **33**:83–100.

Suurmond, R.T., and Bergkvist, E., 1996, *Spectral Pattern Recognition with Neural Networks*, Working Paper, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Vapnik, V.N., 2000, *Statistical Learning Theory*, John Wiley and Sons, New York, NY, USA.

Wierzbicki, A.P., Makowski, M., and Wessels, J., eds, 2000, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Chapter 2

Adaptive Management for Resilience in Human and Natural Systems

Jan Sendzimir

Abstract

Resource management problems have so often defied prediction that surprise rather than certainty has become a common theme for practitioners (managers) and theoreticians. Our understanding of this surprise has improved with our appreciation of resilience and the scales of ecological processes and landscape pattern. But how can we practically address this uncertainty while protecting biodiversity and resilience? I describe a process, Adaptive Environmental Assessment (AEA), that has developed over 30 years of experiments as a test of our abilities to integrate inquiry, understanding, and action in the face of surprising shifts in evolving resource systems. AEA has been applied to resource management problems such as tourism, fisheries, forestry, mining, and agriculture. I discuss current experiments with AEA in North America at large scales (Everglades, Florida Bay) and small scales (dairy farm in Minnesota).

Keywords: Adaptive management, resilience, biodiversity, stability domains, spatial hierarchy, grazing.

2.1 Introduction

The speed and extent of change in natural and human systems are accelerating at unprecedented scales, forcing managers to make a qualitative leap and look over their conceptual horizon to find the sources of change. The qualitative difference in our appreciation of change is more than multi-disciplinary or multi-sectoral; it challenges the foundations of most models of the world as a continuum of various attributes. The qualitative leaps needed to understand the new dimensions of change seem to reflect a hierarchical world in which a few sets of processes control operation and structure over limited ranges of scale. If change is not occurring uniformly everywhere, but only over specific ranges of scale, then understanding must jump from the local to the regional and global strata of the world hierarchy. Our failure to appreciate hierarchy is often compounded by ignorance of the unexpected and non-linear dynamism of human and natural systems. Profound surprise and uncertainty are the result, and they are replacing stability and predictability as the common themes to managing change.

The degree and quality of uncertainty inherent in the dynamics of ecological, social, and economic change can be classified as statistical uncertainty, model uncertainty, or fundamental uncertainty (Hilborn, 1987). Lay discourse about change may acknowledge the shallowest level of uncertainty, statistical uncertainty, wherein one may not know the condition of a variable at any one point, but the overall chances of its occurrence (probability distribution) are known. An example of this might be the chances of being struck by lightning. More profound kinds of uncertainty are currently encountered at the frontiers of science and practice. For example, the depth of surprises occurring in natural and human systems are forcing us to reexamine our most basic ideas about how variables are connected in a model (model uncertainty) or whether we can conceive of any model at all that applies (fundamental uncertainty) (Peterson *et al*., 1997). In the case of model uncertainty one still can predict outcomes but have no idea of their likelihood. For instance, evidence for periodic drops in Europe's temperatures is best explained at present by the switching off of a deep ocean current, the Atlantic Conveyor, yet we have little idea what processes combine to toggle these systems on and off and less of an idea of their likelihood (Broecker, 1996). Fundamental uncertainty applies to situations so novel that no current model applies. The discovery of the atmospheric ozone hole exemplified such profound novelty; we couldn't even bring up a cast of characters let alone a set of relationships between them. One begins to appreciate the complexity of systems when one realizes that, as our Earth is increasingly connected by ecological and human processes, all three levels of uncertainty can apply at any one place. Uncertainty challenges more than our need to understand, because the responsibility to manage human and natural systems creates a tension between the need for useful simplifications that allow discussion (theory) and the need for effective action (practice). This tension increases as the uncertainty springing from Nature is compounded by that contributed by society's attempts to learn and manage. Both natural and human systems are constantly changing and evolving, sometimes in synchrony and sometimes not. If our appreciation of uncertainty in the face of evolution forces us to admit that there are no "truths" which persist, and that no person or group is the guardian of such truths, then we can recognize the importance of discussion between a variety of competing ideas. This raises a serious question: if we admit that we cannot eliminate uncertainty, then what means are available to reduce it when we try to understand and manage unpredictable disruptions?

In this chapter I discuss new theory and practice for understanding and managing uncertainty in systems that incorporate both humanity and nature. I confront two basic questions, "What factors maintain the integrity of these systems?" I will answer this briefly by describing new advances in the theory of ecological resilience. The second question is: "What are useful tools to understand and promote resilience?" I will first discuss briefly some of the sources of uncertainty in nature and society, then I will introduce a process of democratic dialogue, Adaptive Environmental Assessment (AEA), that attempts to practically address the tension between theory and practice by deepening understanding even as the system is managed. I will conclude by suggesting ways AEA could be applied to enhance the understanding and management of floods.

2.1.1 Sources of uncertainty in nature and society

Natural Systems

The unpredictable behavior and surprisingly stratified ('hierarchical') structure of natural systems contribute greatly to uncertainty. Natural systems rarely remain on a constant, predictable course; their behavior can erupt in episodes of transformation, recognized in antiquity in biblical terms: plagues, pestilence, fire, and flood (Holling *et al*., 1995). Forests may appear to grow at a reassuring pace for decades only to be consumed in outbreaks of insect pests or fire. Rare events, such as storms, floods, or biological invasions, can radically and unpredictably restructure systems with effects lasting for long periods. For example, the U.S. Army Corps of Engineers will not guarantee the flow of the Mississippi River through the city of New Orleans, because it has been finally recognized that no practicable level of engineering can prevent certain hurricanes from redirecting the Mississippi down the Achafalaya basin. Such infrequent episodes can also cause systems to jump irreversibly to new states; forests become grasslands, grasslands become shrublands or deserts.

Surprise from natural systems comes partly from our failure to recognize the hierarchical pattern of their behavior and structure. Briefly, ecosystems are not uniform or continuous in space or time, an assumption about pattern that has made predictions much easier to make in the past, but has led to tragic and unforeseen consequences. Natural systems are patchy and heterogeneous in space and discontinuous in time. Forests are not uniform mono-cultures but mosaics of patches of different trees and groups of trees. The processes that give these systems their architecture or structure do not operate uniformly at the same time and space scales. They have different "footprints" because they function at radically different rates and over vastly different spatial extents, often differing by orders of magnitude in time (seconds to millennia) and space (centimeters to kilometers). For example at micro- scales the competition for sunlight and water and nutrients results in plant architecture and operates over square meters in spurts of seconds to hours. Medium scale processes (fire, pest outbreaks, and flood) create and maintain the patchwork of the landscape, operating over square kilometers in episodes that occur every 10 to 50 years. And macro-scale processes, such as geomorphology, structure the landscape over hundreds of kilometers, returning periodically over millennia. Therefore, each stratum (range of scales) in the landscape hierarchy is dominated by a different set of processes; no process is dominant at all scales.

Figure 2.1 shows such a discontinuous world by diagraming the space and time dimensions of different elements of a forest and climate hierarchy. Each polygon shows the minimum resolution (left for space or bottom for time) at which the phenomenon is perceivable, and the horizon (right for space and top for time) over which the phenomenon is replaced. For example, a forest stand is visible on a screen with pixels 10 meters on a side, and most stands are less than 5 kilometers in extent. Similarly, forest stand dynamics can be captured at a minimum time step of a year and a time horizon of a century. These polygons attempt to map out the dimensions at which the processes that create forest stands (or any other element in the hierarchy) operate. In a sense, each polygon is a "footprint" in space and time of the set of processes that dominate at that scale. This diagram pictures the hypothesis that there is no overlap between the scale ranges at which different sets of processes dominate. Sunlight may be omnipresent, but the process of competition for energy, nutrients, and water that result in a plant do not dominate at the scales of kilometers. At that scale, processes such as fire, flood, human agriculture, and forestry dominate to give the meso-scale patterns of the landscape mosaic. Like a Chinese puzzle, the domain of micro-scale processes fits within that of the meso-scale, which in turn fits within that of the macro-scale.

What are the consequences of such a novel world that is not continuous in its behavior or its appearance? These disjunctions in space and time force us to radically revise how we build our understanding up to predict what will happen in

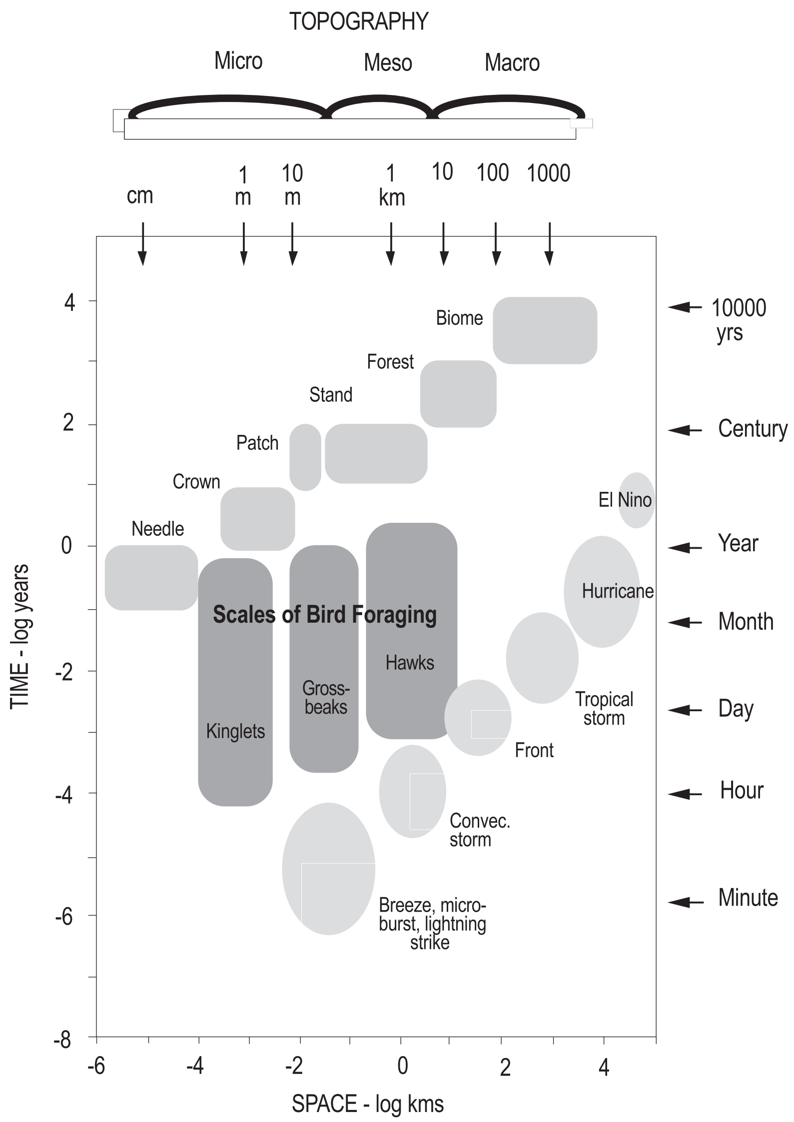


Figure 2.1. Model of discontinuous distribution of space/time dimensions for operation of atmospheric processes, forest structures, and bird foraging scales. Source: Sendzimir *et al*. (1999).

systems as large as nature. They mean that traditional methods of extrapolating from the small to the large, from the present into the future, do not work. Namely, one cannot extrapolate understanding of microscopic phenomena (that which we can most easily observe and test) and scale it up to understand the functioning of the environment at larger scales (forests, towns, regions, states). The local control offered by one dam gives little power to predict the behavior of water over an entire river basin. We must observe and test the processes and phenomena at the appropriate scale, and at larger scales experimental replication and control are often not practicable or possible.

Systems do not remain the same but shift or jump between states. Systems that from a human bias appear stable actually are changing slowly within some limited domain of behavior. Leaps to new domains are the surprises that embarrass theorists and managers. We now recognize from such reversible and irreversible jumps that systems do not have one single balance point or equilibrium. They are often multi-equilibrial, and jumps between different states are increasingly recognized (Holling *et al*., 1995) for their contributions to diversity, structure, and resilience of these systems. What have been labeled as 'disturbances,' with the connotation of degradation from an ideal state, are now seen more as 'invigorating' gymnastics that bolster the long-term integrity of the system. These new insights do not disparage the concept of stability as some source of unhealthy stasis; stability is recognized for its contributions to productivity and bio-geochemical cycles. Therefore, it is not disturbance or stability but the cycling between them that now appears to be the engine of evolution and resilience.

Human Systems

Like natural systems, human systems are also moving targets that occasionally jump erratically in shifting between system types. The uncertainty inherent in shifting natural systems can be amplified by interactions with dynamic human societies that are also disjunct in geographical distribution and behavior. Many societies have moved forward in leaps in terms of technology and/or social institutions, and attempts to understand and cope with nature's variability have quite often built up from initial success to catastrophic collapses. For example, early harvests in some fisheries spurred successive bursts in capital and technology that eventually ratcheted harvest efforts up and fish stocks down to levels requiring possibly a century for recovery (Walters, 1986). Below I briefly discuss how our confidence in dealing with natural catastrophes has been eroded by the mixed success of some institutions and facets of society.

Government, commerce, and science are three broad vehicles for managing uncertainty inherent in complex human and/or natural systems. The constraint of law, the discipline of the market, and the scientific method are all means which partly serve to minimize variability of certain behaviors of people and/or natural resources, or to control the supply and flow of money that tracks these behaviors. The mounting scope of resource management failures has caused widespread loss of confidence in these institutions, both individually and in concert. Governmental failures to understand or manage resources have emerged most strikingly in command-and-control approaches of centralized authority. Such approaches ignore further experimentation or local wisdom as they lock in to one most efficient means of production, and often continue to roll forward on political momentum long after local economies and ecologies have been devastated. The Soviet management of Eastern Europe is one of the most extreme examples of central control resulting in some of the most patent failures to understand or respond to evolving ecosystems or societies. However, non-socialist examples abound because authority is often concentrated in industry and/or government. And the current trend toward globalization of economies can be criticized as an unhealthy concentration of power whose attempts to minimize variability at global levels makes the system more brittle and vulnerable to collapse at world scales.

Sometimes governments and private industry work as partners to try and guarantee smooth and steady economies by suppressing variability and uncertainty of natural variables. Predictable availability of electricity or transport is created by steadying river flow with dams, and dependable deliveries of food result from pesticide use to eliminate sudden outbreaks of insects or microbes. Many of these dual efforts have resulted in massive failures of such shared resources as fisheries, farms, and forestry, or in catastrophic releases of toxic materials. Often government and/or industry have distorted science through clumsy attempts at information manipulation in order to cover the fact that management actions have no real basis in knowledge. Management agencies often suppress scientific dissent in order to present a unified, "certain" front to the outside world, thereby consolidating the political power of the agency (Walters, 1997).

For many, science has lost the aura of a compelling tool for understanding or prediction for a number of reasons. The fact that the same data can legitimately be interpreted in radically different ways is at first baffling and then increasingly ridiculous to the popular mind. One might expect the confusion over science to increase as the scale of disturbances increases, because science loses the ability to replicate and control experiments as their scale expands. While this is true, in addition science suffers from a reputation inflated by revisionist histories that filter out the original controversies surrounding scientific discoveries. In a sense, science is falling from a pedestal created by idealized visions of a history of "strong" science, replete with clean breakthroughs that could relieve us of confusion and uncertainty by dramatic and unassailable demonstrations of causation. Actually, such demonstrations are very rare, and the actual importance of many famous discoveries is only recognized in hindsight. Rutherford's dramatic 1920 "vindication" of Einstein's theory of relativity was actually not a very clear demonstration at all, and was challenged for years by other interpretations (Collins and Pinch 1993). The problem for science as a tool for exploring uncertainty is that few but scientists have the tools, the discipline, or the patience to wade through the controversy and see the real and compelling patterns of evidence emerge over years. And as larger economic/ecological experiments occur in the biosphere, the increasing number of interrelated causes will not clarify the picture sooner; rather, the signals and evidence found will be murkier than before.

The challenge of usefully applying science emerges clearly in some attempts to understand and manage complex systems by quantifying indices of system "integrity." These attempts assume that complex systems are composed of components with relatively constant and tight relationships that consistently behave in a certain way, and, hence, have a 'normal' state against which to compare transient states. Actually, such systems are "open, loosely defined assemblages with only weak evolutionary relationship to one another" (Levin, 1992) and their constant change makes it very hard to define what 'normal' is (De Leo and Levin, 1997). Consistent local disturbance (tidal flux) may allow highly competing species to coexist, or catastrophes (fire, floods) may periodically reset the clock by eliminating most species. While separating the effects of human from natural disturbance is difficult, these problems are compounded by the variety of connections between different components resulting in different functions. Therefore, what 'health' an index reveals is related to which components and which functions are present and measurable at that point in the cycle of change in the system. Quantification may give one a 'spurious sense of certainty' because components have been reduced to numbers and are more easily communicated so as to make a convincing scientific or political statement. As DeLeo and Levin (1997) conclude:

> A more promising approach to ecosystem management is to recognize that various genetic, competitive, and behavioral processes (rather than states) are responsible for maintaining the key features of observed ecosystems, and that the dynamics of these processes vary with the scale of description.

2.1.2 Discontinuous world models and ecological resilience

Beginning in the 1970s, attempts to understand and manage natural resource crises have generated new conceptual models to try to understand why we are so often surprised by natural catastrophes (Holling, 1986). Assumptions of continuity in system behavior and in spatial distribution of resources seemed to blind people to the possibilities of sudden change, so new conceptual models were developed that focus on non-linearities in space and time. Catastrophe theory (Casti, 1982) emerged to explore non-linear system dynamics and hierarchy theory (Allen and Starr, 1981; O'Neill *et al.*, 1986).

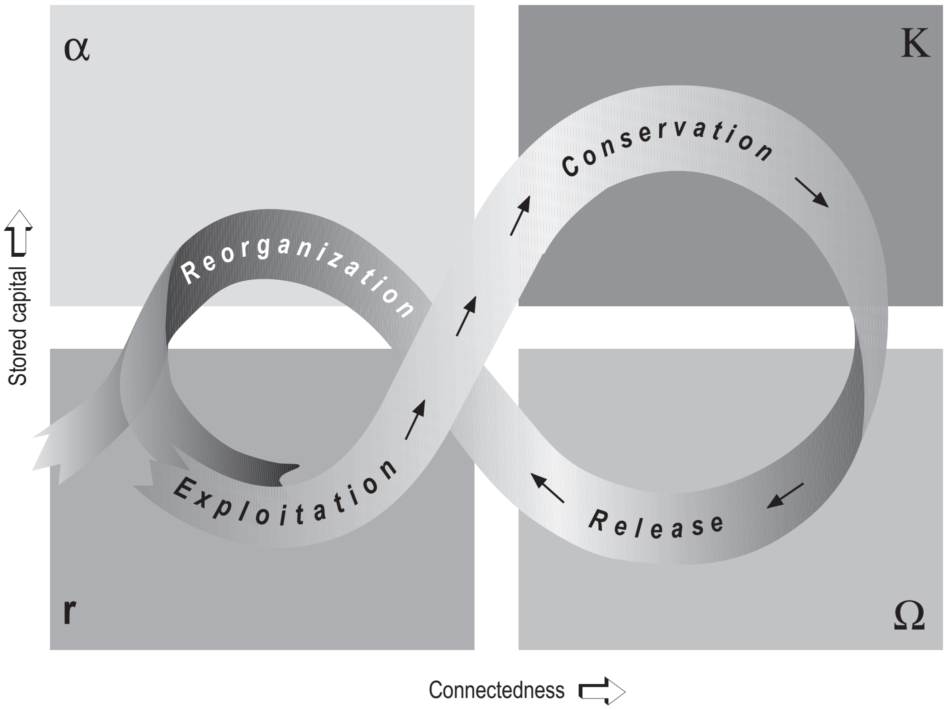


Figure 2.2. The Adaptive Cycle – The four phases of ecosystem dynamics that correspond to the functions of exploitation (r), conservation (K), release (Ω), and reorganization (α). Source: After Holling *et al*. (1995).

Holling (1973, 1992, 1995, 1996) integrated these new models of a hierarchical world to develop "ecological resilience" as an overarching concept of the functional relations that sustain the integrity of systems. He illustrates system dynamics using the Adaptive Cycle (*Figure 2.2*) to portray sudden change as inevitable, emerging from the endogenous dynamics of the system, not as an inexplicable departure from the norm created by exogenous factors. This cycle divides system dynamics into four phases, commonly viewed from the "birth" of a system as it self-organizes from a relatively undifferentiated state. The phases are: Exploitation (r), Conservation (K), Release (Ω), and Reorganization (α). The first two phases (r to K) superficially resemble the classical Clementsian view of ecological succession from barren ground to climax forest, which makes any sudden change in the system's trajectory look like a "disturbance" that prevents the system from realizing an "ideal" end-state. Holling (1996) uses the Adaptive Cycle to extend the Clementsian view to incorporate surprising deviations, catastrophe, and renewal. The transition from r to K shows how self-organization enhances the system's stocks to the point where it eventually becomes so dense and over-connected that it is "an accident waiting to

Table 2.1. Factors that contribute to the resilience of human and natural systems.

Control of disturbance	Regulation of renewal or regenerative potential
Disturbance frequency and intensity	Stored resources
– Chesapeake shellfish fishery	– Soil depth, organic content, seed bank
– Herbivore grazing/browsing	– Water (aquifer, lake, river)
– Fire in forests, grasslands	– Nutrients in biomass
– Lightning in mangroves	
Capacity to absorb disturbance	Facility of response
– Landscape morphometry	– Re-colonization distance
– Habitat availability	– Biodiversity
– Ability to migrate	– Cross-scale functional reinforcement
(connectivity of landscape)	– Within-scale functional diversity
Processing and cycling of resources	Availability of information
– Cross-scale functional reinforcement	Viability of cultural information transfer language
– Within-scale functional diversity	Customs (education, discourse)
	Politics
	Human memory
	Population age structure

happen." At that point, any contagious process (fire or pest outbreaks) can spread a pandemic of destruction (Omega phase) that releases the system's resources. The future of the system resides in how these resources are recaptured and used to build a new system. This pivotal juncture, when a forest may degrade to a grassland or desert, or a lake may suddenly shift from clear water to an algal broth, is represented by the Reorganization (Alpha) phase.

The Adaptive Cycle illustrates the potential paths of change as a series of dynamic transitions that can renew systems periodically when their resilience is high or can degrade them when their resilience declines. Resilience has no numeric measure. It is a qualitative indicator of a system's capacity to maintain its integrity. It focuses on how much shock or change a system can undergo and still remain the same system. For example, the rich, productive grassland of the Jornada valley supported intense grazing for centuries in New Mexico. Within just a few years in the late 19th century the grassland shifted to a shrub desert unfit for grazing, though no major change in farming practices had occurred. The system's resilience declined to the point where almost any small factor could cause the entire system to flip to another state.

While resilience theory has not advanced to the point of quantifiable indices, it usefully focuses attention on those factors that sustain and promote resilience.

Two broad categories of factors that contribute to resilience are control of disturbance and regulation of renewal (*Table 2.1*). In the first category, while disturbances are inevitable, their effect may be less than catastrophic. Communities will probably survive and thrive on those disturbances with frequencies and intensities to which they have evolved for long periods of time. Experimentation continues to improve management practices (controlled fire or grazing) that can effectively achieve proper disturbance rates and intensities. Disturbance intensity can also be adjusted technologically to achieve viable economic/ecological systems. For example, the Chesapeake Bay shellfish fishery was headed for extinction due to overexploitation until the state government set a technical limit on fishing capacity by requiring that all fishing vessels be powered by sail. This lowered the fishing disturbance intensity to a level that allowed for viable shellfish populations.

Resilience can also be maintained by factors that increase a system's capacity to absorb disturbances. For example, river basin landscapes with their original (unchannelized) morphometry have a wider cross-section and can absorb higher flood volumes. Dutch water managers are now starting a program to abandon dikes and channels and reinstate the floodplain morphometries that were originally created and shaped by flooding events (Middelkoop and de Boo, 1999). Another spatial factor that contributes to resilience is the configuration of habitats in the landscape, but the contribution is not always positive. Highly fragmented landscapes are more resistant to invasions and to contagious spread of disease but their lack of connectivity may also lead to collapse of animal populations that need mobility to find resources or to reproduce.

System resilience is sustained as well by a diverse and redundant capacity to process energy, nutrients, and resources. Peterson *et al*. (1998) have integrated wildlife ecology with hierarchy theory by suggesting that terrestrial animals perceive a discontinuous landscape and exploit only limited ranges of scale (strata or levels within the landscape hierarchy). Tiny birds search for insects at the finest landscape level, the leaves and needles in the trees, whereas larger birds search for insects or rodents at much coarser levels, such as fields and river edges. Because different animals use a diversity of resources within each scale range Peterson *et al*. (1998) propose that ecological function is redundant within each geographic scale range. For example, within a single scale range, such as a tree canopy, different animals use a variety of resources, consuming different groups of insects, fungi, vegetation, mammals or birds.

However, such diversity and redundancy of function exists not only within a single scale range but across all scale ranges as well. Different animals exploit the same resources but at different levels in the hierarchy, so ecological function is redundant across all scale ranges within the landscape. For example, tiny birds may seek and eat individual caterpillars on a single tree branch, but larger birds will come and pursue the same prey when a caterpillar population explosion causes them to saturate an entire patch of trees with high densities. A caterpillar population explosion makes itself evident at the next larger scale, the tree patch. Therefore, at different times, different kinds of birds exploit the same resource (caterpillars) at micro as well as meso-scales. The resilience of such a system is sustained by this capacity to utilize resources and keep them cycling at different times and different scale ranges. A system that loses such capacity will accumulate resources in ways that invite new species to invade and exploit them or new processes to emerge. For example, fire may become more important if biomass begins to accumulate. In this way a system can shift its character, changing the communities of plants and animals that inhabit it.

The other category of factors (*Table 2.1*) that enhances resilience is "Regulation of Renewal." Once a system's resources are released in the destructive Ω phase, what factors exist that allow the system to retain those resources and to reorganize and re-establish itself? Stored resources (soils, seed banks, water, and nutrients) certainly retard resource dispersal, and/or contribute stored resources that promote production and pull loose resources into living biomass. Some factors facilitate the response function of resource rescue and renewal. For example, recolonization by plants or animals will be aided if seeds or animals have short distances to travel to disturbed zones. These recolonization distances are shortened by the landscape's spatial distribution and diversity of habitats. The potential for redevelopment is also enhanced by the same redundancy of function within and across scales previously discussed. Biodiversity contributes to that potential by providing a variety of species which utilize resources at different scales of space and time.

Renewal and regeneration are also promoted when the system can reliably find and use information about the system's history. Information can be stored in language, custom, literature, educational tools and traditions, political processes, and human memory. This alludes to the hypothesis that human survival was greatly enhanced once our genes and/or our customs promoted the survival of people old enough to remember long-term events, crises such as floods, fires, droughts, and plagues. A population age structure with sufficient elderly members also has greater reproductive potential among a variety of animals (fish, mammals, birds).

The significance of resilience theory and the indicators it suggests is that it allows one to appreciate the complexity of system dynamics and spatial heterogeneity and yet concentrate on the critical factors (turning points, the spatial patterns) that are functionally related to system collapse or renewal. It does not eliminate uncertainty; nothing does. But it provides concepts and vocabulary that help narrow uncertainty to a workable level on which new theory and practice can be tested even as a complex system is managed.

If our initial successes in eliminating variability and uncertainty have led to more profound catastrophes, how can we responsibly engage or embrace uncertainty and effectively respond to change? The challenge for society is that not only must understanding be consistently pursued and deepened to appreciate dynamic and evolving systems, but that one must take action in the midst of this effort. In other words, coping with novelty and surprise requires the sustained capacity to learn and to flexibly manage. For thirty years a decision making process has been evolving to address the twin challenges of learning and management. This process, Adaptive Environmental Assessment (AEA), has been refined in a series of on-the-ground applications in problems of forestry, fisheries, national parks, and river systems. It is currently being applied in two North American river systems, the Mississippi and the Colorado, and offers opportunities to address the development of society on flooding riparian systems. I will describe with examples some of the theory and operation of the AEA process.

2.2 Adaptive Management

2.2.1 Underlying assumptions

As previously discussed, the driving assumption underlying AEA is that uncertainty is inevitable, because the behavior of natural resource systems is only partly knowable. Therefore, as ecosystems and societies evolve, so humans must adapt and conform as systems change. However, the challenge of environmental problems denies us the luxury to constrain our focus simply to understanding. Society must respond at a number of levels that include both understanding and management. Historically, the understanding that was developed in isolation from the discipline of reacting to and managing a changing system has often proven shallow and of limited use. Therefore, AEA is not about learning before one can manage, rather it is learning while one manages (Gunderson, 1998).

How can management and learning be coordinated? Based on the assumption that structured learning is better than trial and error, AEA is based on a process of Integrated Learning (*Figure 2.3*). As Gunderson (1998) notes, “The process is structured for learning by systematically probing uncertainties of resource issues, continually assessing, postulating, testing and re-evaluating.”

If evolving complex adaptive systems are fountains of uncertainty, and surprise is inevitable, then structured learning is the way that uncertainty is winnowed. Surprise is never eliminated, but we may reduce the consequences of the way our understanding lags behind evolving systems by embracing uncertainty, deepening understanding, and adaptively responding to system changes. Adaptive responses and management actions must meet social objectives, such as protecting people or

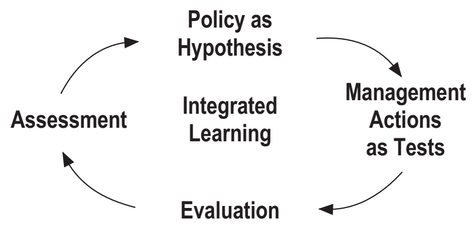


Figure 2.3. Key ingredients contributing to structured learning in the AEA process.

resources, but learning must continue as policies are modified to adapt to surprises. And therefore, a second function of management is to probe the system, perturbing it slightly to provoke some minimal, safe response that gives an indication of the working and true structure of the system (Walters, 1986). In this way, AEA views policies as hypotheses, therefore management actions become treatments in an experiment.

I shall now discuss in turn the functioning of the different phases of AEA, how uncertainty is confronted by formulating hypotheses, how management actions test these hypotheses, and how learning integrates assessment and management. I shall then describe one example of AEA as applied in a wetland savanna ecosystem in Florida.

2.2.2 Assessing the known and the uncertain

The assessment phase simultaneously engages two apparent opposites, integrated understanding and uncertainty, and counter-poses them in ways that are revealing to both. Rather than dodging uncertainty with simplifying assumptions or rationalizations, the AEA process focuses on uncertainty from the very beginning, utilizing disagreements to reveal and highlight gaps in understanding and other sources of uncertainty. The adaptive process identifies new bases for sharing understanding when gaps or uncertainties are recognized as common to all the different disciplines, sectors, occupations, trainings, and experiences represented in the discussion.

The common gaps and links in understanding can bridge the various backgrounds present and establish a foundation of trust that may eventually unlock information and experiences that were previously unshared. This trust is one way in which the AEA process addresses the refusal to share information, a frequent source of gridlock in environmental decision processes. Another way is to select representatives of various backgrounds based on competence, respect within their group, and the willingness to cooperate. Participants are given to understand, that a great potential for communication can emerge if only each person “leaves his/her gun at the door," be that gun an opinion, a philosophy, or a mandate from one's organization.

The assessment phase aims to initiate and foster discussion by using an informal workshop setting and computer models. Care is taken to introduce and use computer models simply as translators and integrators of people's understanding, not as technically superior vehicles of "truth." If dialogue begins where there was none before, then the computer model has succeeded. If people begin to seriously reassess their assumptions because model output based on their ideas seems questionable, then important and novel insights are possible.

The goal of the assessment phase is to integrate understanding and ponder uncertainties to the point that they can be clearly stated as hypotheses about how the system works and what effects interventions (management or uncontrolled human actions) might produce. Complexity in adaptive systems is partly the result of the diversity of causes, and the alternative explanations that address these causes can become the basis for policy in the next phase.

2.2.3 Policies as hypotheses

Policies are the governing plan, the question set based on experience that sets the stage for further action. Policies range from the formal (government acts, laws, administrative code, legal contracts) to the informal (understandings and shared views among groups). Instead of pursuing the 'correct' policy as a solution to problems, AEA differs from traditional engines of policy by looking for policy that addresses other social objectives as well as the need to learn in the face of uncertainty (Gunderson, 1998). In this light, policies are not magic bullets that address the right mix of objectives to solve a problem, rather they are astute hypotheses about how the world works or "Questions masquerading as answers" in the words of Steve Light. AEA embraces uncertainty by trying to find the best questions, and thereby tries to dodge the trap of assuming certainty by rallying around 'solutions.'

2.2.4 Management actions as tests

Many environmental problems stem from administrative pathologies that narrow policy to achieve efficiency at the expense of awareness about where the system is going. For example, if initial policies achieve high production, one could bank on maximizing the profit of such success by cutting research costs, but only if one was sure of where the system is going. The AEA process strives to avoid this pathology by broadening implementation to mean the testing and evaluating of hypotheses (policies). This prevents the intent of policy from being changed during implementation, and shifts the search for efficiency from cost reduction to checking whether management actions were executed as anticipated (Gunderson, 1998). This gives implementation a disciplinary rigor of consistency in execution, because otherwise the test of the policy becomes meaningless, and one has lost the power to gain new information about the system.

2.2.5 Integrative learning

Amassing information does little to help anticipate surprise and uncertainty. Projections based on previous system behaviors have limited utility in the face of true novelty. Integration of the information gained in policy probes has little to do with data quantity and everything to do with quality. To what extent have we winnowed uncertainty and closed the gap on these elusive and dynamic systems? Enhancing understanding through integrated learning is a second loop type of learning that is fundamental to adaptive management in several ways. First, it integrates across multiple disciplines and backgrounds. Second, the focus group, and the community at large, learns by doing. All this deepens understanding by probing the workings of ecosystems and society and by applying the considered and thoughtful sharing of new ideas and previous experiences. Such inquiry is structured by expert facilitation of discussion, which sums up new insights and consolidates gains before reformulating the questions at hand. Finally, this understanding often builds from ground made more fertile by complete re-inspection of assumptions and conceptual frameworks (Gunderson, 1998).

2.2.6 The Everglades: An example of AEA applied

One of the key objectives of adaptive assessments of a resource issue is to highlight uncertainties and generate a number of plausible hypotheses about the issue. The AEA process develops these hypotheses as a suite of alternative explanations about the behavior of the resource. The process of considering the suite of competing ideas helps to integrate concepts about ecology, economy, or politics and to weigh the various policy options. Therefore, the hypotheses link our understanding of the issue with the range of possible outcomes that management actions might produce (Gunderson, 1998). I illustrate this below using the example of wading bird declines in a wet savanna known as Everglades National Park in Florida.

Wading bird populations have declined dramatically (as much as 95 percent) over the past 70 years in South Florida (Bancroft, 1989). The Everglades National Park provided a primary nesting site for millions of birds at the beginning of this century, and these numbers have declined to the tens of thousands. During an AEA process convened in 1989, a number of alternative hypotheses were posed to explain these population declines (Light *et al*., 1995). I briefly paraphrase each alternative explanation below.

Shrunken Habitat: The conversion of portions of the Everglades by agriculture and urbanization has decreased the original area to half its size. This area has low biological productivity per unit area, so loss of productive habitat has led to lower nesting populations.

Decreased Flow: The development of the Everglades involved drainage and diversion of much of the water in south Florida to the extent that much less water flows through the park. These lower water flows have caused dramatic declines in biological productivity at the estuarine fringe of mangroves, a border area that used to hold the densest nesting colonies.

Damped Fluctuations of Water Level: Water levels fluctuate seasonally in South Florida, driving the ecology of the Everglades. These fluctuations provide the means of food production and delivery. Fish populations thrive and reproduce in times of flooding and are concentrated by lowering water levels to the point where wading birds can easily feed on them. Water management schedules for canals in the Everglades have changed these hydrological patterns to the point where they are not synchronized with wading bird nesting cycles.

Distant Magnet: The decreases in nesting populations in the Everglades are matched by increases in other parts of the southeastern United States: Louisiana and the Carolinas, for example. Population declines in the Everglades may not wholly reflect lowered ecological conditions there so much as better or improving conditions elsewhere that have drawn the populations to distant sites.

Mercury: Mercury concentrations have increased in the atmosphere over this century, and many wetland soils absorb and concentrate deposition from the air. Anaerobic water conditions can mobilize this metal from the soil, and it can pass up the food chain to wading birds. Over time the latent toxic effects of mercury have decreased the nesting success of wading birds.

Parasites: Increased agriculture upstream of the Everglades has released progressively larger amounts of nutrients into the surface water, and populations of parasites have thrived and increased as a result. The increased burden of parasites has diverted metabolic energy normally given to reproduction and thereby lowered the success of nesting of wading birds.

2.2.7 Passive and active adaptation

How can understanding of these alternative explanations be integrated at the same time that one must manage the system? Walters (1986) introduced three concepts of how to structure management approaches in the AEA process:

1. Evolutionary ("trial and error"), which starts with a haphazard set of choices and progressively winnows these down to a better subset to improve results;
2. Passive Adaptive, which applies historical data to select or construct a response model ("single best estimate"), with the management decision being made assuming this model is correct; and
3. Active Adaptive, which uses historical data to establish a suite of competing hypotheses or response models, and the manager's policy choice reflects a balancing of anticipated performance in the short term with the longer term advantage of knowing which hypothesis is most correct (Walters and Holling, 1990).

Two problems arise with passive adaptive approaches. First, the effects of management interventions are confounded by effects of the environment. This is evident in the long and bitter debates about whether fishing effort or environmental effects (climate, watershed habitats lost to silt from logging) are primarily to blame for collapsed fisheries (Walters and Collie, 1988). A second, and more fundamental, problem is that passive adaptive policies may allow us to miss opportunities to improve the system's performance. This might occur if the 'right' model and the 'wrong' model both predict the same response pattern, and the system is managed as if the wrong model is correct (Walters and Holling, 1990).

So what should a manager do in pursuing an active adaptive approach so as to properly engage a suite of alternative explanations? No hypothesis has an exclusive lock on the truth, and each is to some degree plausible. The answer lies in balancing between two areas:

1. Considering the policy implications of the entire suite of hypotheses and
2. Developing a process to sort between all the hypotheses (Gunderson, 1998).

In the first case, if all hypotheses point toward similar policies, then one can proceed and manage in a flexible way. In the Everglades example above, if all hypotheses pointed toward water dynamics as the reason for nesting loss, then a set of management experiments could be developed to test these ideas. One set of tests would address most or all hypotheses at the same time. If the suite of hypotheses do not point toward the same policy implications, then any policy that is firmly and irreversibly established would be doomed from the outset. For example, if the Distant Magnet hypothesis were closest to the truth, then any water-based policy would not only fail to achieve the conservation goal but would erode the trust of stakeholders who are participating in the AEA process (Light *et al.*, 1995).

The second approach, sorting between competing hypotheses, is generally done in the assessment stage of AEA. In the case of the Everglades, an active adaptive approach might have recommended a policy of monitoring wading bird populations at much larger scales while experimenting with a qualitatively different set of manipulations (water flow, periodicity, or nutrient removal) to try to tease out which of the competing explanations holds the most promise. The AEA process counters the tradition of casting a policy into concrete through law by iteratively testing these sets of hypotheses through the years and making recommendations to adapt as results and understanding develop.

2.3 Adaptive Processes Applied to Overgrazing

Adaptive management can successfully be applied at scales as small as a single farm. AEA has been used as a framework for effective collaboration between scientists, farmers, and citizens in exploring new agricultural practices that mimic ecological functions.

In the early 1990s six dairy farmers in southeastern Minnesota began experimenting with new ways to feed their cattle out of concern for higher commodity prices and the effects of overgrazing.[1] They dropped conventional cropping to explore rotational grazing, an approach that relies on the farmer to move grazers in response to changes in indicators of ecosystem health. A farm is subdivided into sections (paddocks), and cattle are moved from paddock to paddock for short periods of intense use followed by long periods of recovery. This idea has many roots, one of which recently began in Africa from observations that wildlife grazing caused less erosion than cattle. This idea grew to practical experiments to mimic with cattle the way wildlife would intensely utilize an area and then move on, giving the area a long rest before returning. These experiments eventually coalesced into an ecological management approach called Holistic Resource Management (HRM) which a local NGO, the Land Stewardship Project, had introduced to the region in a series of workshops. However, some Minnesota farmers had developed forms of rotational grazing on their own in decades past, so exploring this idea represents either a leap back in time (before the jump in agriculture intensity of the 1970s) or in space (to modern wildlife ecology emerging from Africa).

The Land Stewardship Project worked to create a partnership with the farmers, local citizens, government environmental agents, and scientists and students at the Minnesota Institute of Sustainable Agriculture. This alliance used an adaptive

[1] King, T., and DeVore, B., Bringing the land back to life, The Sierra Club, http://www.sc.org/sierra/199901/goodfarms.html

framework to develop hypotheses about what were indicators of the crucial ecological, economic, and social processes on the farm. They then worked jointly to monitor experiments with different grazing patterns (frequencies in time and distributions in space), modifying experiments and indicators as their understanding changed.

The results summarized in the list below show a broad range of benefits ecologically, economically, and socially. The key lesson for scientists is that even promising new theory and practice may take 10 years or more of experimenting to become practical in a particular ecosystem or society. But this coalition successfully applied ideas about African wildlife ecology on another continent. They showed that cattle could be an ecological and economic benefit if the cattle were managed to mimic the disturbance pattern (in space and time) that the system had evolved with, probably with buffalo. And the new ideas gained public support as the experience of participating citizens spread informally through society. In summary, the experiment advanced scientific theory and practice at the same time that it strengthened the rural social network and the economies of the farms.

The accomplishments of this Adaptive Management experiment in southeast Minnesota are shown in *Table 2.2*.

Such experiments are instructive in how to develop programs that are practical in how one defines and probes to achieve what is "natural." Definitions of what is natural can confound science and management when they are arbitrary and have little relationship to the operation of ecological processes. For example, "natural" is often defined in the United States as the state of ecosystems prior to contact with Europeans. However, Botkin (1990) notes that ecosystems have changed dramatically in species composition and spatial patterns for many millenia before humans arrived in North America. There is no one ecosystem state that is the "original" or "natural" one; nature is a moving target. Similarly, Vera (1999) has shown through pollen analysis of lake bottom sediments that climax vegetation in Central and Western Europe in prehistoric times was not closed forest but more open and savanna-like due to herbivore browsing. Therefore, current management of parks as closed forest ecosystems may be based on an artificial, human misconception of what is "natural." Restoring the importance of ecological processes (such as grazing) rather than species lists (biodiversity) to the definition of "natural" would help in correcting this misconception. And it requires sustained, flexible cooperation between scientists and non-scientists to experiment and discover the dimensions of ecological processes that make it resilient, and therefore, sustain its "naturalness." The same can be said for economic, and social processes. So the advantage offered by Adaptive Management is a rigorous scientific framework for experimenting with processes (ecological, economic, and social) that sustain the resilience of systems (both human and natural). Experiments are currently underway in Poland to see

Table 2.2. Accomplishments of the Adaptive Management experiment on southeast Minnesota farms.

- The farms are successful at a time when 30 dairy farms a day fail.
- Biodiversity has soared to 100 bird species on farms with no pesticides.
- The farmers have re-established their own social institutions - local networks of inquiry, knowledge, and encouragement among themselves and in partnership with local citizens, government employees, and academics.
- Knowledge is being passed on as the next generation apprentices on these farms and as other farmers and citizens use the adaptive methods developed here, now available on video as The Monitoring Toolbox.
- Farmer insights pushed ecological and agricultural science such that more respectful working relations between farmers and scientists bode well for more productive future collaborations.
- Farmers were enabled to take their risky insights all the way to proven agricultural production systems once they had the backing and trust of a partnership of NGOs, government, and academia. But these were the innovative farmers who need ideas less than they need the security of funding and trust to try their insights. This project does not address the needs of less innovative farmers.
- Knowledge and respect for farming and science are percolating through rural communities as people discuss their participation in monitoring over the dinner table and in the living rooms.
- The study exploded the myth of farming as a crippling disturbance. Stream erosion was severe in the total absence of disturbance (no cows) or if cows were allowed to visit the stream anytime. The farmers tinkered until they found the correct rate of disturbance (cow visits to the stream) and then erosion was minimized.
- Good ideas rarely work off the shelf. Farmer insights took a decade of experimentation before the better practices became clear. This highlights the value of long term support for long range collaborations between farmers, NGOs, and scientists.

what level of herbivore grazing is not a disturbance but a boost to the biodiversity and resilience of floodplain ecosystems in the Narew valley.

2.4 Conclusions

The policy-based experimentation advocated by adaptive management is essential to reduce the ecological, social, and economic costs of learning. Adaptive management focuses upon developing alternative hypotheses, identifying gaps in knowledge, and assessing what knowledge would most effectively distinguish alternative hypotheses and, therefore, could be most useful in setting and updating research and action priorities. As Peterson *et al.* (1997) state:

> Rather than simply testing and rejecting individual hypotheses, scientists and decision makers must consider diverse sets of alternative hypotheses. Alternatives need to be continually revised, modified, and discarded, based upon how they fare in tests against empirical data (Hilborn and Mangel, 1997). Maintaining the status quo must be explicitly examined as one alternative among many, with its attendant consequences, benefits, and costs. More often than not, policy decisions have multiple dimensions that are difficult, if not impossible, to convert into a single metric. In these cases, techniques such as multi-attribute utility analysis, wherein tradeoffs between alternatives are evaluated using multiple metrics, may be necessary. In either case, such methods of analysis are best viewed not as authoritative objective procedures, but as modeling processes that provide a means of making underlying valuations open to scrutiny, discussion, and sensitivity analysis.

In order to exercise reasonable caution we should recognize that the greater our uncertainty, and therefore the less our capacity to precisely define risk, the more considered and "reversible" our management actions should be. Data accumulation and analysis may narrow our sense of uncertainty, but our capacity to predict risk is persistently undercut by the scale of our actions in creating new uncertainties. Adaptive processes provide one of the most prudent frameworks for assessing and addressing the multiple scales at which flooding risk and damage emerge.

The laboratory for the theory and practice about floods has to be wider even than society; it has to span the range from local village experience to global sources of weather processes. The hard lessons of the last 40 years mandate that we learn to address all these scales, flexibly and repetitively, so that the most important question is always at hand.

References

Allen, T.F.H., and Starr, T.B., 1982, *Hierarchy: Perspectives for Ecological Complexity*, University of Chicago Press, Chicago, IL, USA.

Bancroft, G.T., 1989, Status and conservation of wading birds in the Everglades, *American Birds*, **43**:1258–1265.

Botkin, D., 1990, *Discordant Harmonies – A New Ecology for the Twenty-First Century*, Oxford University Press, Oxford, UK.

Broecker, W.S., 1996, Thermohaline circulation, the Achilles heel of our climate system: Will man-made CO2 upset the current balance?, *Science*, **278**:1582–1588.

Casti, J., 1982, Catastrophes, control and the inevitability of spruce budworm outbreaks, *Ecological Modelling*, **14**:293–300.

Collins, H., and Pinch, T., 1993, *The Golem: What You Should Know About Science*, Cambridge University Press, Cambridge, UK.

De Leo, G.A., and Levin, S., 1997, The multifaceted aspects of ecosystem integrity, *Conservation Ecology [online]*, **1**(1):3, available from the Internet, http://www.consecol.org/vol1/iss1/art3.

Gunderson, L., 1998, A Primer On Adaptive Environmental Management, Arthur R. Marshall Laboratory, Department of Zoology, University of Florida, Gainesville, FL, USA.

Hilborn, R., and Mangel, M., 1997, *The Ecological Detective: Confronting Models with Data*, Monographs in Population Biology, Princeton University Press, Princeton, NJ, USA.

Hilborn, R., 1987, Living with uncertainty in resource management, *North American Journal of Fisheries Management*, **7**:1-5.

Holling, C.S., Peterson, G., Marples, P., Sendzimir, J.P., Redford, K., Gunderson, L., and Lambert, D., 1996, Self-organization in ecosystems: lumpy geometries, periodicities and morphologies, in *Global Change and Terrestrial Ecosystems*, Cambridge University Press, Melbourne, Australia.

Holling, C.S., Schindler, D.W., Walker, B.W., and Roughgarden, J., 1995, Biodiversity in the functioning of ecosystems: an ecological synthesis, in C. Perrings, K-G. Maler, C. Folke, C.S. Holling, and B-O. Jansson, eds, *Biodiversity Loss: Economic and Ecological Issues*, Cambridge University Press, Cambridge, UK.

Holling, C.S., 1992, Cross-scale morphology, geometry and dynamics of ecosystems, *Ecological Monographs*, **62**(4):447–502.

Holling, C.S., 1986, The resilience of terrestrial ecosystems; local surprise and global change, in W.C. Clark and R.E. Munn, eds, *Sustainable Development of the Biosphere*, Cambridge University Press, Cambridge, UK.

Holling, C.S., 1973, Resilience and stability of ecological systems, *Annual Review of Ecology and Systematics*, **4**:1–23.

Interagency Floodplain Management Review Committee, 1994, Sharing the Challenge: Floodplain Management into the 21st Century, Report of the Committee to the Administration Floodplain Management Task Force, Washington, DC, USA, June 1994.

Levin, S.A., 1992, Orchestrating environmental research assessment, *Ecological Applications*, **2**(2):103–106.

Light, S., Gunderson, L., and Holling, C.S., 1995, The Everglades: Evolution of management in a turbulent ecosystem, in L. Gunderson, C.S. Holling, and S. Light, eds, *Barriers and Bridges to the Renewal of Ecosystems and Institutions,* Columbia University Press, New York, NY, USA.

Middelkoop, H., and de Boo, eds, 1999, Twice a River; Rhine and Meuse in the Netherlands, RIZA Report, Ministry of Transport, Public Works and Water Management, pp. 127, ISBN 90 369 52 239.

O'Neill, R.V., DeAngelis, D.L., Waide, J.B., and Allen, T.F.H., 1986, *A Hierarchical Concept of Ecosystems*, Princeton University Press, Princeton, NJ, USA.

Peterson, G., Allen, C., and Holling, C.S., 1998, Ecological resilience, biodiversity and scale, *Ecosystems*, **1**:6–18.

Peterson, G., De Leo, G.A., Hellmann, J.J., Janssen, M.A., Kinzig, A., Malcolm, J.R., O'Brien, K.L., Pope, S.E., Rothman, D.S., Shevliakova, E., and Tinch, R.R.T., 1997, Uncertainty, climate change, and adaptive management, *Conservation Ecology [online]*, **1**(2):4, available from the Internet, http://www.consecol.org/vol1/iss2/art4

Sendzimir, J., Light, S., and Szymanowska, K., 1999, Adaptively understanding and managing for floods, *Environments*, **27**:115–136.

Vera, F.W.M., 1999, Metaforen voor de wildernis. Eik, hazelaar, rund en paard (trans. Metaphors for the wilderness; Oak, hazel, cattle and horse), Dutch Ministry of Agriculture, Nature Management and Fisheries, ISBN: 90-5485-746-3.

Walters, C.J., 1986, *Adaptive Management of Renewable Resources*, McMillan, New York, NY, USA.

Walters, C., 1997, Challenges in adaptive management of riparian and coastal ecosystems, *Conservation Ecology [online]*, **1**(2):1, available from the Internet, http://www.consecol.org/vol1/iss2/art1.

Walters, C.J., and Collie, J.S., 1988, Is research on environmental effects on recruitment worthwhile?, *Canadian Journal of Fisheries and Aquatic Sciences*, **45**:1848–1854.

Walters, C., and Holling, C.S., 1990, Large-scale management and learning by doing, *Ecology*, **71**(6):2060–2068.

Chapter 3

Modeling Techniques for Complex Environmental Problems

Marek Makowski

Abstract

Mathematical models can be useful in decision-making processes whenever the amount of data and relations are too complex to be analyzed based solely on experience and/or intuition. Models, when properly developed and maintained, and equipped with proper tools for their analysis can integrate relevant knowledge available from various disciplines and sources. Most environmental decision problems are complex. However, some of them pose additional challenges owing to the large amount of data, the complex relations between variables, the characteristics of the resulting mathematical programming problems, and the requirements for comprehensive problem analyses. Such challenges call for applications of advanced techniques for model generation and analysis. Several of these techniques are outlined in this chapter and illustrated by the RAINS model, a large non-linear model, which has been used in international negotiations about the reduction of air pollution.

Keywords: Modeling paradigms, decision support systems, air quality, object-oriented programming, robustness, multicriteria model analysis, non-linear optimization, model management.

3.1 Introduction

Most decision problems are no longer well-structured problems that are easy to solve by intuition or experience supported by relatively simple calculations. Even the same kind of problem that was once easy to define and solve, has now become much more complex because of the globalization of the economy, and a much greater awareness of its linkages with various environmental, social, and political issues. Modern decision makers (DMs) typically want to integrate knowledge quickly and reliably from these various areas of science and practice. Unfortunately, the culture, language, and tools developed to represent knowledge in the key areas (e.g., economy, engineering, finance, environment management, social and political sciences) are very different. Everyone who has ever worked on a team with researchers and practitioners having backgrounds in different areas knows this. Given the great heterogeneity of knowledge representations in various disciplines, and the fast-growing amount of knowledge in most areas, we need to find a way to integrate knowledge for decision support efficiently.

Rational decision making is becoming more and more difficult, despite the quick development of methodology for decision support and an even quicker development of computing hardware, networks, and software. Two commonly known observations support this statement:

- first, the complexity of problems for which decisions are made grows even faster;
- second, knowledge and experiences related to rational decision making develop rapidly but heterogeneously, therefore integration of various methodologies and tools is practically impossible.

A critical element of model-based decision support is a mathematical model, which represents data and relations that are too complex to be adequately analyzed based solely on experience and/or intuition of a DM or his/her advisors. Models, when properly developed and maintained, can represent not only a part of knowledge of a DM but also integrate relevant knowledge available from various disciplines and sources. Moreover, models, if properly analyzed, can help the DM to extend his/her knowledge and intuition. However, models can also mislead users by providing wrong or inadequate information. Such misinformation can result not only from flaws or mistakes in model specification and/or implementation, the data used, or unreliable elements of software, but also by misunderstandings between model users and developers about underlying assumptions, limitations of applied methods of model analysis, and differences in interpretation of results, to name a few. Therefore, the quality of the entire modeling cycle determines to a large extent the quality of the decision-making process for any complex decision problem.

A recent comprehensive overview of model-based decision support methodologies, tools, and environmental applications is provided in Wierzbicki *et al*. (2000). The monograph[1] also contains a detailed discussion on the modern decision making process, and on guidelines for model development and analysis, focusing mainly on multicriteria model analysis (MCMA).

This chapter concentrates on an overview of modeling paradigms and techniques applicable to complex models and illustrates them by the RAINS model. The structure of the chapter is as follows. The RAINS model is outlined in Section 3.2, which is followed by a discussion of modeling problems and applied techniques in Section 3.3. Section 3.4 presents an overview of MCMA methods.

3.2 Outline of the RAINS Model

In many parts of Europe the indicators of critical levels of air pollution are exceeded and measures to improve air quality in these areas are needed to protect the relevant ecosystems. Several international agreements have been reached over the last decade in Europe to reduce emissions. For several years, the Transboundary Air Pollution (TAP) Project[2] at IIASA has been developing models that have been used for supporting international negotiations. The models help to identify cost-effective measures aimed at reducing various measures of ground-level ozone concentrations, acidification, and eutrophication at several hundred receptors over Europe. These measures correspond to policies for reducing emissions of NH_3 (ammonia), SO_x (sulphur oxides), NO_x (nitrogen oxides), and VOC (volatile organic compounds) by various economic sectors in European countries.

The structure of the RAINS model is outlined in *Figure 3.1*. The decision variables are composed of the levels of emissions NH_3, SO_x, NO_x, and VOC by various sectors in each country, which imply the corresponding emission control policies. For each country and type of emission, a cost function is defined. Such a function relates the emission level with the corresponding costs of reducing to this level a sum of various types of emissions caused by activities aggregated (for the purpose of this analysis) for each country in several sectors. Therefore, cost-effective measures can be calculated by minimizing the cost function that corresponds to the sum of costs related to reductions of all types of considered emissions in all sectors in each country. In order to determine the corresponding environmental impact, emission levels are used as inputs to the three dispersion submodels and to the ozone formation submodel. Studies of the impact of ozone, acidification, and eutrophication have resulted in the establishment of critical levels for various air-quality indicators in order to protect agricultural crops and forests. These are

[1] http://www.iiasa.ac.at/∼marek/pubs

[2] http://www.iiasa.ac.at/Research/TAP

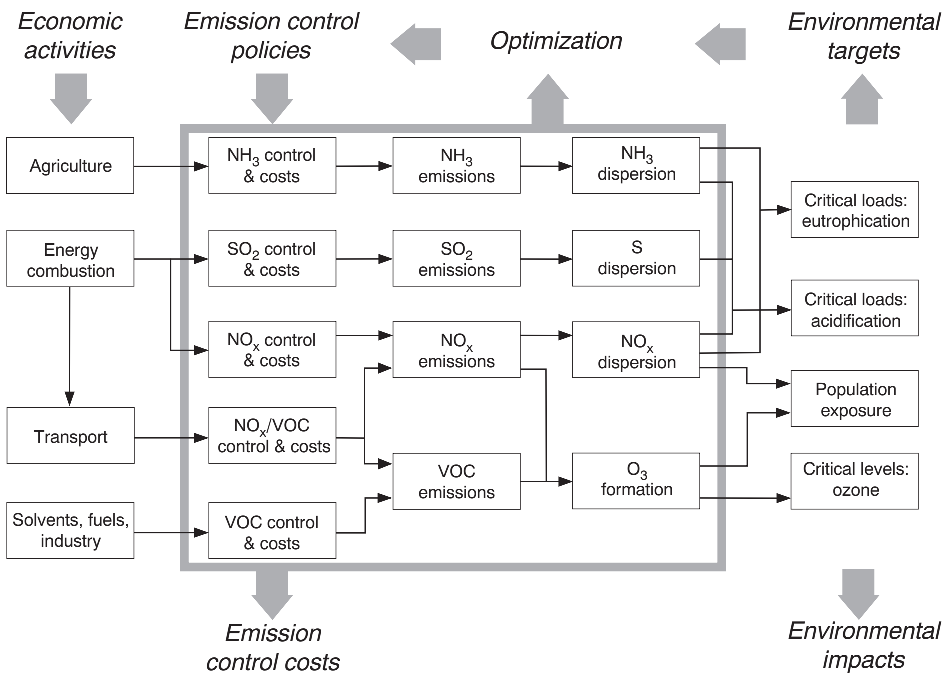


Figure 3.1. RAINS model structure.

determined using a long-term exposure measure, called the *accumulated excess*. Consequently, nine such exposure indices (six for ozone, two for acidification and one for eutrophication) have been defined for each of approximately 600 grids in Europe (also called *receptors*), and accumulated excess PWL (piece-wise linear functions) are defined for each grid and for each type of acidification and eutrophication excess. These indices are used to assess environmental effects of the applied emission control policies.

It is not only the structure of the RAINS model that is complex, but also the way in which it is used. In 1989, when the sulphur protocol was due for renegotiation, the United Nations Economic Commission for Europe (UN-ECE) accepted the RAINS model for use in the negotiations. Most probably, this was the first time when all parties to a major international negotiation accepted one computer model as a key tool in their negotiations. However, the acceptance of the model was just the beginning. The negotiators had to trust the model results and understand how the model works. The scientists had to understand the political realities and modify the model in order to respond better to the requests of the negotiators. In order to illustrate just one element of this process, let's consider an interpretation of the optimality of a solution. From the scientific perspective, a rational optimality criterion is a minimization of the sum of costs of emission reductions subject to constraints on values of the air quality indices. However, this obviously results in solutions that would oblige some countries and/or industries to make larger emission reductions than others (which also implies substantial costs). Acceptance of such a solution would certainly distort competition; therefore, negotiators cannot accept such solutions. On the other hand, the RAINS model clearly demonstrates that uniform reductions (which are a sound idea from a political point of view) would not only be much more expensive but also would not result in achieving the desired air quality. Another example of this mutual learning process undertaken by the negotiators and scientists is illustrated by the evolution of the understanding of what the desired air quality should be. For example, the results of extensive research have shown that the critical acid loads should vary substantially between various ecosystems. Therefore, there is no justification to apply uniform environmental requirements for all grids in Europe.

In mathematical programming terms, RAINS is a large (about 30,000 variables and over 30,000 constraints), non-linear model. The original RAINS model described in Alcamo *et al*. (1990), which was a small linear programming (LP) model that dealt only with acidification, can be considered as a small pilot prototype of the current version of RAINS described in this chapter. The development of several versions of RAINS, made over ten years, was driven by the needs of the negotiators. The first version of RAINS was used for negotiating the sulphur protocol; therefore, it dealt only with a single pollutant. However, it has become clear that a multi-pollutant, multi-effect approach offers substantial environmental and financial advantages. Therefore, to respond to these needs, RAINS has been extended and gradually modified to its current form. A description of the current version of RAINS and of its use can be found in Schöpp *et al*. (1999), while a more formal model specification and a more detailed discussion of applied modeling paradigms is provided by Makowski (2000).

3.3 Modeling Problems and Techniques

Modeling of any complex problem is composed of the following mutually linked activities:

- model specification,
- data handling and model generation,
- model analysis.

These issues will be discussed in the following subsections.

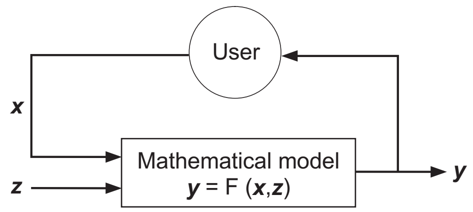


Figure 3.2. A mathematical model represents relations between decisions (inputs) $\boldsymbol{x}$, external decisions (inputs not controlled by the user) $\boldsymbol{z}$, and consequences (outcomes) $\boldsymbol{y}$.

3.3.1 Model specification

Mathematical models are widely used in many areas of science and industry for predicting the behavior of a system under particular circumstances, when it is undesirable or impossible to experiment with the system itself. The understanding of the system gained through a comprehensive examination of its model can greatly help in finding decisions, the implementation of which will result in a desired behavior of the system. Therefore, a model used for decision support is focused on the basic function of a DSS (Decision Support System), namely, to provide an evaluation of consequences that will result from an implementation of given decisions.

All four of the basic concepts illustrated in *Figure 3.2*, namely, decision variables, external decisions, outcome variables, and a mathematical model are briefly discussed in the following subsections.

Decision Variables

In model-based decision support it is assumed that decisions have quantitative characters and therefore can be represented by a set of the model variables, hereinafter referred to as decisions[3] $x \in E_x$, where E_x denotes a space of decisions. In a trivial case $x \in R$, which denotes that a decision is represented by a real number. However, in most cases x is a vector composed of various types of variables. For larger problems, the components of x are grouped in several subvectors. Let us illustrate this by specification of the decision variables of our illustrative model.

In the RAINS model the main decision variables are the annual emissions of the following four types of primary air pollutants from either sectors or countries:

- n_{is}, annual emission of NO_x from sector is;
- v_{is}, annual emission of nonmethane VOC from sector is;

[3]For the sake of brevity we call decision variables simply decisions.

- a_i, annual emission of NH_3 from country i; and
- s_i, annual emission of SO_2 from country i.

where vectors n_{is} and v_{is} are combined for each country in subvectors n_i and v_i, respectively.

Additionally, optional decision variables are considered for scenarios that allow for limited violations of air quality targets. For such scenarios, variables corresponding to each type of considered air quality target are defined for each receptor. Each variable represents a violation of a given environmental standard. Optionally, violations of targets can be balanced with surpluses (understood as the difference between a target and its corresponding actual concentration/deposition).

External Decisions

Figure 3.2 illustrates two types of inputs to the core model: (1) decision variables $\boldsymbol{x}$ controlled by a user, and (2) external decisions denoted by $\boldsymbol{z}$. In practice, inputs $\boldsymbol{z}$ may include representations of various quantities that substantially influence the values of outcomes $\boldsymbol{y}$ but are not controlled by the user, for example:

- regulations or commitments on environmental standards for air or water quality management models;
- meteorological conditions assumed for modeling physical relations in environmental models, e.g., *average, wet, dry, worst* year data for a water model; or
- forecasts of changes in demand for services, e.g., in telecommunication or transportation models.

In the RAINS model the external decisions $\boldsymbol{z}$ are represented by:

- values representing the environmental standards that define constraints for various indices (such as maximum concentrations of various water and air quality indicators, respectively); and
- the set of meteorological data used for calibration of a respective model.

While the external decisions are beyond the control of the user of a DSS, s/he typically wants to examine a range of scenarios with various representations of external decisions in order to find out not only a solution which will best respond to a most likely representation of external inputs $\boldsymbol{z}$, but also a solution that will be *robust*, i.e., will also be good for various compositions of $\boldsymbol{z}$ that should be considered.

Outcome Variables

The consequences of implementing various decisions $\boldsymbol{x}$ are evaluated by values of outcome variables $y \in E_y$. In various fields of applications, outcome variables are named differently, e.g., outcomes, metrics, goals, objectives, performance indices, attributes.

In the RAINS model, one outcome variable represents the sum of costs of reductions of emissions; four sets of other outcome variables correspond to various indices of air quality. While the definition of the cost is rather simple, an appropriate definition of air quality indices is rather complex. Environmental effects caused by acid deposition, by excess nitrogen deposition (the latter defined for two types of critical loads), and by eutrophication are evaluated at each receptor by a PWL function that represents an accumulated excess over the threshold of the environmental long-term target. If optional violations of environmental standards are allowed, then a maximum (over a set of receptors in each country) violation of each type of air quality indicator is also considered as an output variable.

Objectives

Out of the set of outcome variables $y \in E_y$, a user selects a subset of variables conventionally called objectives $q \in E_q$, where E_q is a space of objectives. Quite often objectives are referred to as criteria, and in this chapter these two terms will be used interchangeably. Usually E_q is a subspace of E_y, that is, the DM selects several criteria q_i from the set of outcomes y_j. Sometimes also some of the decision variables x are used as criteria, but for the sake of consistency we assume that such a variable is simply represented by one of the outcomes y.

The difference between objectives and outcome variables is not strict, and is mainly determined by the preferences of users. It has been commonly observed that a human being typically prefers to deal with seven plus/minus two criteria at a time. However, a complex model usually has many outcome variables. While users of models typically concentrate analysis by specifying preferences for several selected objectives, values of all outcome variables are reported, and sometimes a modification of the selection of objectives is desired. Therefore, depending on the stage and type of model analysis, the selection of the set of objectives is modified.

A partial preordering in E_q is usually implied by the decision problem and has obvious interpretations, such as the minimization of costs competing with the minimization of pollution. However, a complete preordering in E_q cannot usually be given within the context of a mathematical programming model. In other words, it is easy to determine for each objective separately, which solution (represented by vectors x and q) is the best one. However, for conflicting objectives there are two sets of solutions:

- Pareto-optimal (often called efficient), i.e., a solution, for which there is no other solution for which at least one criterion has a better value while values of remaining criteria are the same or better;
- dominated, i.e., solutions that are not Pareto-optimal.

Obviously, a Pareto-optimal solution is preferred over any solution it dominates (assuming that the selected criteria correspond well to the preferential structure of a DM). However, a set of Pareto-optimal solutions (often called Pareto-set, or Pareto frontier) is typically composed of an infinite number of solutions, many of which are very different. Pareto-optimal solutions are not comparable in a mathematical programming sense, i.e., one can not formally decide which is better than another one.

However, DMs are able to express their own preferences for various efficient solutions. One of the basic functions of multiobjective decision support is to provide various ways in which a DM may specify his/her preferences. There is no reliable formal way for separating a specification of preferences from a process of learning from the model analysis. It is a commonly known fact that decision making is not a point event, even in situations where it is realistic to assume that the problem perception does not change during the decision-making process. Therefore, the possibility of using a DSS in a learning and adaptive mode is a critical feature.

Mathematical Model

As already illustrated in *Figure 3.2*, a mathematical model (further on also called a core model) is used for predicting the consequences of decisions x, which can be either proposed by a DM or computed by a DSS. The consequences are measured by values of outcome variables y. Therefore, a model can be represented by mapping $y = F(x, z)$, where $x \in E_x$, $z \in E_z$, and $y \in E_y$ are vectors of values of decisions, external decisions, and outcomes, respectively. For the sake of brevity we will assume further on that the external decisions z are given and represented as parameters of the mapping F.

The core model (often called also substantive model) should include only logical and physical relations that are necessary to adequately represent relations between inputs x and outputs y. In addition to inputs and outputs, a model contains various intermediate and parametric variables (balance and/or state variables, resources, external decisions), conventionally called auxiliary variables. In a typical complex model, the decision and outcome variables are a small fraction of all variables. Auxiliary variables are introduced for easing the model specification and handling, and are typically not interesting for an end-user of the model. However, the way in which auxiliary variables are selected and defined has a critical impact on the model performance and reliability.

In other words, the core model is composed of decision, outcome, and auxiliary variables, and of relations (inequalities, equations, etc., conventionally called *constraints*) between these variables that indirectly determine the sets of admissible (feasible) decisions and the corresponding solutions. Some of the constraints may reflect the logic of handling events represented by variables. For example, the model known as RWQM (Regional Water Quality Model) (Makowski and Somlyódy, 2000) has the constraint:

$$\sum_{k \in K(j)} x_{jk} = 1 \qquad x_{jk} \in \{0,1\}, \qquad j \in E \tag{3.1}$$

where $K(j)$ is the set of technologies considered for emission node j, and E is the set of nodes where emissions occur. This condition assures that exactly one technology (represented by the corresponding binary variable x_{jk}) is selected in each waste water treatment plant.

Generally, the core model implicitly defines a set of feasible decisions $X_0 \subseteq E_x$. In other words, x is feasible, if and only if $x \in X_0$. The set X_0 is composed of all vectors x that fulfill all constraints representing all logical and physical relations among all the variables used in the model. Since every actual (and properly defined) decision problem has at least two solutions, X_0 is not empty.

A reader familiar with mathematical programming may be surprised, that such a model does not contain any goal function. This is done on purpose, and it is the recommended way of implementing any model-based DSS. We shall explain now, why the core model should not contain any representation of a preferential structure of a DM.

It is usually not possible to specify uniquely a model that can yield a unique solution reflecting the preferences of a DM. For example, very often it is practically impossible (even for a good analyst or an experienced DM) to specify values for a group of constraints that would determine a solution that corresponds well to preferences of a DM. In order to illustrate this point let us consider the RWQM model. A DM typically considers different wastewater treatment technologies and the related costs, as well as standards for water quality. However, s/he knows that specification of constraints for a group of (either ambient or effluent) water standards may lead to solutions that are too expensive. On the other hand, assuming constraints for costs (with water quality standards being goals) could result in an unacceptable water quality. Values of constraints are in such cases formally parameters in a corresponding optimization problem. But those values are, in fact, decisions that reflect the preference structure of a user. Setting constraints' value too tight would result in restricting the analysis of the problem to a (possibly small) part of feasible solutions (often making the set X_0 empty). Textbooks on modeling typically provide the advice of using sensitivity analysis to deal with these limitations. However, as discussed in the section below on sensitivity analysis (see Section 3.3.3), applicability of sensitivity analysis to complex problems is very limited. A more practical approach in such situations is to specify two types of constraints, so called hard and soft constraints which correspond to *must* and *should* types of conditions, respectively. But, in fact, dealing with soft constraints can easily be done within multiobjective model analysis, which will be discussed later.

Therefore, the specification of a core model that defines X_0 should not include any relations that reflect conditions for acceptability of a solution by a user or a preferential structure of a DM. Hence, the *core model* accounts only for logical and physical relations between all the variables that define the set X_0 of feasible solutions. All other constraints and conditions that implicitly define acceptability of a solution by a user and those that represent a preferential structure of a DM will be included into an interactive procedure of the model analysis. This provides the flexibility of examining trade-offs between various solutions.

Such an approach to model specification and analysis allows us to design and implement a model-based DSS, which is conceptually composed of two parts:

- A constant and usually large core model. This part is built and verified before an actual analysis of a problem starts.
- A part that corresponds to a current specification of preferences defined by a user. This specification is interactively being changed, often drastically, by a DM.

Proper implementation of such an approach makes it possible for a DM to analyze feasible solutions that best correspond to his/her preference structure. Changing this structure is the essence of the model analysis and of the model-based decision support. There is an additional bonus in the fact that there always exists a feasible solution of the underlying mathematical programming problem, which is a prerequisite for an analysis of complex models.

Finally, we should point out that the value of a mathematical model as a decision aid comes from its ability to adequately represent reality. Therefore, there is always a trade-off between the requested accuracy (realism) of the model and the costs (also time) involved in developing it and providing the model with data. Hence, the requested accuracy should be consistent with the accuracy really needed for the model and with the quality of the available data.

Specification of the RAINS Mathematical Model

We shall now briefly comment on the specification of the RAINS model, which can be considered (as illustrated in *Figure 3.1*) as composed of three mutually linked parts:

- emission control costs and resultant emissions,
- atmospheric dispersion and tropospheric ozone formation models,
- environmental impacts.

Each of these components is backed up with a large amount of underlying research, which is presented in various specialized publications, see, e.g., Schöpp *et al*. (1999) and the RAINS Web site.[4]

Here we can provide only a general overview of these components.

The emission-cost module consists of three parts, estimating current and future levels of emissions of NH_3, SO_x, NO_x, and VOC from each considered sector. These estimates are based on national statistics and projections of economic activities taking into account implemented and possible emission control measures and associated costs. These data are used to define parameters of PWL functions, which map for each sector considered in each country the emission levels of each type of pollutant to the corresponding cost.

The atmospheric dispersion processes over Europe for NH_3, SO_x, NO_x, and VOC compounds are modeled using results of the European EMEP[5] model, developed at the Norwegian Meteorological Institute and described, e.g., in Olendrzyński *et al*. (2000). However, the EMEP model is far too complex to be used for optimization, or even for many scenario analyses. Therefore, an essential requirement of an integrated assessment of the RAINS model is a simplified but reliable description of the dispersion processes in order to represent the source-receptor relationships involved. It is possible to envisage several ways of condensing the results of more complex models to achieve this. One approach is to use statistical techniques to build a simplified model based on the results obtained from a complex mathematical model for a large number of emission reduction scenarios. Such an approach has been implemented for, and is currently used by, the RAINS model. Of course, using simplified source-receptor relationships between the precursor emissions and the various thresholds of corresponding levels/loads results in a lesser accuracy than that assured by the EMEP photo-oxidants model. Therefore, selected results obtained from the simplified model are compared with results from

[4]http://www.iiasa.ac.at/∼rains

[5]EMEP: European Monitoring and Evaluation Programme, cooperative program for monitoring and evaluation of the long-range transmission of air pollutants in Europe (see www.emep.int).

the EMEP model. This is done by running the EMEP model for the emissions obtained from the RAINS model, and comparing the levels/loads values provided by both models.

Another approach, which focuses on specification of a simplified ozone formation submodel, is based on using fuzzy-rules generation methodology and is presented by Ryoke *et al*. (2000). This method uses fuzzy rule generation methodology to represent numerous results of the EMEP model as a response surface describing the source-receptor relationships between ozone precursor emissions and daily tropospheric ozone concentrations. It has been shown that the fuzzy model provides better predictions of ozone concentrations than the traditional regression model based on all data at each grid. Furthermore, the membership functions (MFs) obtained appear to be sensible. When meteorological data are examined, the different fuzzy rules describe different meteorological conditions rather well. Unfortunately, the development of a fuzzy model requires manual tuning of parameters for each grid, therefore the model has been developed only for several grids in Europe. For these grids, the fuzzy model can be used to examine daily ozone concentrations caused by a selected emission scenario in a much faster and easier way than can be accomplished by the much more detailed EMEP model.

Space limitations prevent a full specification of the RAINS model. Such a specification is presented by Amann and Makowski (2000). Here we only outline two issues of more general interest, which are accounted for in this model specification: (1) the optimization problem and (2) soft constraints.

1. The resulting optimization problem has practically non-unique solutions. More exactly, it has many very different solutions with almost the same value of the original goal function. Let's consider two solutions x_1 and x_2 such that:

$$|c(x_1) - c(x_2)| < \epsilon \qquad ||x_1 - x_2|| > \delta \tag{3.2}$$

   where $c(\cdot)$ is a goal function, $||\cdot||$ is a norm used for determining the distance between vectors x_1 and x_2, and ϵ, δ are two positive numbers, small and large, respectively. For most large optimization problems, this is a typical issue that, unfortunately, does not attract enough attention because analysts often look only at an optimal solution without analyzing other solutions, which have practically the same value of the goal function. Typically, a problem gets noticed when various instances of the mathematical programming problem that differ very little have very different optimal solutions (while the goal function remains practically the same).

   There is a simple and practical technique called regularization, which provides a suboptimal solution that has additional properties specified by a user. The methodological background of regularization is presented, e.g., by

Makowski (1994a), and its implementation in the RAINS model is discussed in Section 3.3.3.

2. Representing environmental targets traditionally via hard constraints would result in the recommendation of expensive solutions. Only a few grids have active constraints for environmental targets, and for almost all grids the actual values of indices are substantially lower than the corresponding targets. In order to provide a more complete analysis, so called *soft constraints* (with compensation for the violation of original targets in some grids by a larger margin in other grids) can optionally be specified for environmental targets. They result in much cheaper solutions with more uniform differences between environmental targets and actual values of corresponding indices. The application of soft constraints in the RAINS model is presented in detail by Amann and Makowski (2000), and the mathematical background and also applications to other environmental problems can be found in several chapters of Wierzbicki *et al.* (2000).

3.3.2 Model generation and data handling

There are basically two approaches to the generation and analysis of a mathematical programming problem: either develop a problem-specific generator or use a modeling system (such as GAMS, AMPL, AIMMS). Several issues should be considered when selecting one of these approaches. These problems are discussed in detail by Makowski (2000). Here just seven of them are outlined:

- *Increasing the Efficiency of Model Generator Development* – A modeling system greatly simplifies the task of model specification, especially if compared with the amount of resources needed for the development of a model generator using traditional procedural programming languages like Fortran or C. However, the use of C++ substantially reduces this difference, especially with the Standard Template Library (recently included in the C++ standard), and with other class libraries supporting the implementation of mathematical programming problems.
- *Processing Input Data and Checking Data Consistency* – A model generator is more efficient in processing the input data needed for model specification. It is also preferred when a more sophisticated check of data consistency is desired.
- *Preprocessing* – A modeling system has limited possibilities for efficient preprocessing of optimization problems. This is not a serious problem for linear models because preprocessing is a standard feature of any good linear program-

ming (LP) solver. However, the preprocessing of non-linear models is much more difficult, as demonstrated, e.g., by Drud (1997).

- *Choosing a Starting Point* – For a large optimization problem, a good starting point might dramatically decrease the computation time. The computation of such a point is much easier for a problem-specific generator.
- *Comprehensive Model Analysis* – A modeling system greatly simplifies model analysis within the paradigm specific to a given system. However, using different modeling paradigms – such as soft and/or inverse simulation, regularization, soft constraints, or MCMA – which is necessary for comprehensive analysis of any complex problem, typically requires much additional effort if the particular paradigm is not included in a given modeling system.
- *Computing Nonlinear Constraints and Jacobian* – A modeling system releases a modeler from the complex task of providing code to compute the values of non-linear constraints and the non-linear elements of the Jacobian. However, a typical non-linear problem has only a few formulas for the non-linear part. Therefore, one can use, e.g., *Mathematica* (Wolfram, 1996) for generating C language code for formulas of the Jacobian and for the values of constraints, and then include this code in a class that provides values for particular elements of the Jacobian and for the constraints.
- *Decreasing Costs for Widely Distributed Models* – Finally, for models that are not only run on various platforms but are also widely distributed, a problem-specific generator substantially decreases costs for the users (typically, the cost of a solver plugged into the problem-specific software is a small fraction of the cost of the run-time license for a modeling system).

Taking into account the above-summarized points, the problem generator of the RAINS model has been implemented as a problem-specific C++ class that uses a template class library supporting the generation of mathematical programming problems. The generator includes an efficient preprocessor, which dramatically reduces the size of the non-linear optimization problem, and also performs an instance-specific scaling, which results in values of the Jacobian and Hessian that are unlikely to cause numerical problems for a non-linear solver.

The approach is conceptually very simple. Each of the above-mentioned solvers is available as a library of Fortran subroutines. The generator has C++ classes that are specific for each solver. These classes are inherited from the base classes that handle a common part of the generator. A problem-specific report writer processes the results into a form that eases their interpretations. Another class supports a portable interface between C++ and Fortran. Hence, three versions of executables can easily be produced, each composed of the generator, preprocessor, postprocessor, and one of the solvers.

A nonlinear solver requires routines that compute values as well as elements of the Jacobian of the non-linear constraints and the goal function. A large part of the total computation time is used for the execution of these functions, and therefore the efficiency of their implementation is important. The code for the Jacobian has been generated by *Mathematica* (Wolfram, 1996) with prior use of the *FullSimplify* operator, which simplifies the formulas substantially. This is an easy way to generate an efficient and bug-free code.

The RAINS model requires processing a large amount of data coming from various sources, including other complex models. Therefore, the data used by the TAP Project is maintained by several database management systems, which are coupled with other applications. To make the handling of data used in the RAINS model efficient and portable, the public domain library Hierarchical Data Format (HDF) (Koziol and Matzke, 1998), developed by the National Center for Supercomputing Applications, Illinois, USA,[6] has been employed. The basic data structures are handled by a collection of well-tested template C++ classes that are also used for the LP-DIT.[7] A C++ interface class has been implemented to make handling of the used data structures by the HDF library easy and efficient.

Costs of emission reductions discussed above are given as PWL (Piece-Wise Linear) functions of the corresponding emission levels. PWL functions are not smooth. Therefore, in order to be able to use efficient nonlinear solvers (which require smooth functions), the PWL cost functions are represented by corresponding smooth functions. Due to space limitations, these conversions are not presented here; however, they are described by Amann and Makowski (2000).

Finally, one should notice that the dimensions of the model are not fixed. For some scenarios, a part of the constraints and/or variables does not need to be generated. Moreover, the dimensions of the matrices and vectors used in the model definition vary substantially for various types of analysis. Fortunately, properly implemented constructors of C++ template classes handle such problems in a natural and efficient way.

3.3.3 Model analysis

There are many approaches to model analysis and typically a problem-specific combination of various approaches is needed for a comprehensive analysis of any complex problem. We summarize some general concepts of model analysis that are of more general interest, and then illustrate the need of combining some of the various methods by outlining the approach applied to the analysis of the RAINS model.

[6] http://hdf.ncsa.uiuc.edu/HDF5

[7] Linear programming data interchange tool.

General Concepts

One typically distinguishes two types of model-analysis methods, which are conventionally called simulation and optimization. They can be characterized as follows:

- In *simulation,* decision variables are inputs and goals are outcomes. Therefore this technique is good for exploring the intuition of a DM, not only for verification of the model, but also for providing a DM with information about the consequences – typically represented by values of outcome variables and constraints – of applying certain decisions. One can also consider simulation as an alternative-focused method of analysis that is oriented toward examining given alternatives.
- *Optimization* can be considered as a goal-oriented (value-focused) approach that is directed toward creating alternatives. Optimization is driven either by formulating a single objective in single-criterion optimization, or several objectives in multicriteria optimization, and looking for values of decision variables that optimize the value of the specified objective(s). Therefore, goals are the driving force and the values of decision variables are the outcomes.

Simulation– and optimization-based approaches are in fact complementary. For simulation, one needs to provide values for all decision variables. For this purpose, one may use random values for variables (as proposed by Goodwin and Wright, 1991, who present various techniques and examples), or assign values based either on the DM's intuition or on a heuristic (possibly based on information from a knowledge base). One should, however, note that applicability of these appealing ideas is limited to rather small models; for models having hundreds or even more variables, a specification of values for all decision variables based on intuition is practically unrealistic.

However, even for a large model, simulation can be useful for a *"what if"* type of analysis, e.g., for comparing the results from optimizations with the outcomes from values of decision variables defined by the user, typically by modification of their values obtained from optimization. Of course, there is no way to assure that a given specification of the values of decision variables will result in a feasible solution. Therefore, instead of using a classical approach to simulation, one should use a *soft simulation*, where setting given values of decision variables $\hat{x}$ is replaced by minimization of an outcome variable defined as:

$$\epsilon\|x - \hat{x}\| \tag{3.3}$$

where ϵ is a given positive number, x is a vector composed of decision variables, and $\hat{x}$ is a vector composed of the corresponding desired values of these variables.

In a most simple soft simulation approach, one sets $\epsilon = 1$ and assigns to $\hat{x}$ the given values of decision variables. However, a similar approach may be used also for more sophisticated types of analysis, where x is composed of not only decision variables, and the choice of $\hat{x}$ depends on the desired properties of the solution. In particular, if values of some elements of $\hat{x}$ are not known, then one can set them to be equal to zero, which implies a preference for the minimum norm solution.

Of course, term (3.3) can be used to define an outcome variable that can be used as a criterion in MCMA. It can be used also as a term in a composite goal function with larger values of the parameter ϵ for various simulation techniques. For example, by using a large value of ϵ (i.e., one that dominates the other terms of the goal function) and setting $\hat{x}$ equal to desired values of decision variables, one can find a solution that is closest to such values. If these values are feasible, then a solution composed of these values will be found. If they are not feasible, then the closest feasible solution will be found. Note that in the latter case a traditional simulation will simply report *"infeasible problem*."

For such an approach to soft simulation, the original goal function takes the role of the regularizing term, while for small values of ϵ, term (3.3) may be used as a regularizing term for the original goal function. An application of such an approach in the RAINS model is discussed in the next subsection.

Sensitivity Analysis

In mathematical programming, sensitivity analysis is typically understood as an analysis of changes of an optimal solution caused by an alteration of the data in the model. A traditional approach to such an analysis is based on properties of an optimal solution. It typically consists of calculations of ranges of changes of parameters for which an optimal solution does not change, and on using a dual solution for calculations of changes in value of a goal function for changes in some parameters that are small enough to allow such a simple evaluation procedure. These methodological topics, which all form the subject of post-optimal analysis, and the corresponding software tools have been extensively developed, especially for LP types of problems. However, their applicability is practically limited to rather small, linear models.

There are several problems concerned with applying the classical approaches to sensitivity analysis to problems represented by complex models. We summarize here only the three most important issues:

- The range of changes of parameters for which the classical sensitivity analysis is valid is typically too small to justify its application to models of mixed-integer and non-linear types.

- The concept and tools for sensitivity analysis have been developed and implemented for analysis of rather small models. Complex models are typically large, however; therefore use of these techniques is either cumbersome or practically impossible for complex models.
- In many models, the quality of dual solution is rather questionable, and for many other models the dual solution is practically non-unique. This is owing to the fact that most large models are numerically badly conditioned, and due to efficient presolve algorithms, which greatly decrease the resources (time and memory) needed for solving large problems. However, presolving always guarantees the quality of the primal solution but often results in an unreliable dual solution, which is the basis for classical sensitivity analysis. Therefore, a reliable sensitivity analysis requires a good understanding of various techniques and corresponding tools, which is rather limited to highly skilled specialists in mathematical programming.

Generally, one distinguishes two groups of problems which correspond to the two related but distinct issues that are typically used for a justification of applications of sensitivity analysis, namely:

- Model development, where some parameters of the model can hardly be precisely determined; here by parameters we understand only coefficients in logical and physical relations.
- Model analysis, where a classical single-objective optimization-based approach forces the analyst to treat all but one actual goal as constraints.

A discussion on how and when the selection of a type of model (such as fuzzy or stochastic) can adequately represent a problem for which a deterministic model with fixed parameters may be too simplified is far beyond the scope of this chapter. In many practical applications, a deterministic model is an adequate simplification provided that the developers of the model have enough data and experience to properly evaluate values of parameters. In some situations, a parametric analysis of a model is nevertheless needed, but this is typically done during the model validation. Another technique that is useful, and is more efficient than some elements of sensitivity analysis, is a specification of so-called *soft constraints*, and the use of such constraints for a definition of outcome variables.

The second issue (model analysis) can be easily addressed by using MCMA, which is based on a core model that does not include the preferential model of a user. In classical single-criterion optimization, several objectives were typically treated as constraints, for which one had to specify an acceptable value. This approach has not only the disadvantages discussed above, but it also requires analysis

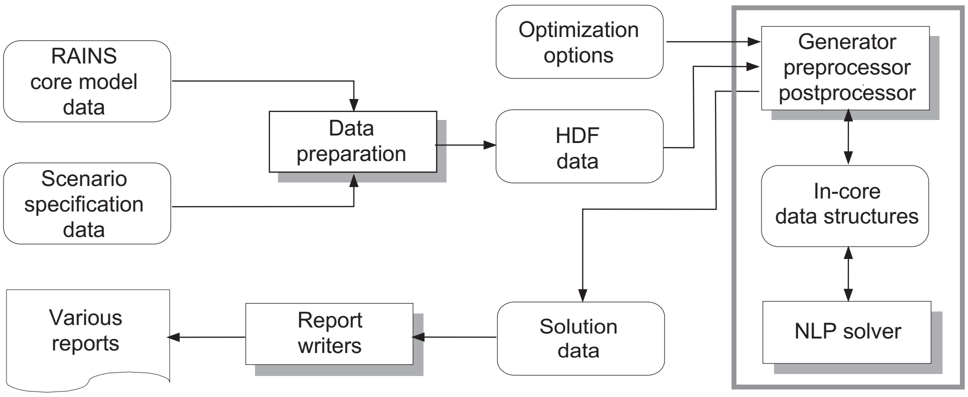


Figure 3.3. RAINS model analysis cycle.

of the impact of changes caused by specified constraining values for criteria that are treated as constraints. Such values cannot, in practice, be specified precisely, therefore their modifications are inevitable. Sensitivity analysis was developed in order to help in analysis of such modifications. However, the functionality of sensitivity analysis, which was applied to this part of classical analysis of optimization models, is replaced in multiobjective model analysis by more robust and natural approaches to problem analysis. Therefore, multiobjective model analysis offers better ways for providing some of the functionalities that are theoretically promised by sensitivity analysis.

Analysis of the RAINS Model

This section outlines how a combination of various methods of model analysis has been applied to the RAINS model, which is used extensively for various types of analysis that are needed for supporting international negotiations. Obviously, neither RAINS nor any other complex model provides any “best” solutions. This is simply because there are several problems and trade-offs that are both moral and social. No model can actually answer such questions, and this remains the domain of negotiations. However, models can help the negotiators concentrate on those parts of the negotiations that should not be represented by a mathematical model. This assistance is provided by various unbiased analyses, such as computing the consequences of given policies of emission reductions, or advising the values of emission levels that correspond best to given criteria and constraints.

Due to space constraints, I have limited this section to presenting the structure of one cycle of analysis followed by a summary of the implementation of a composite goal function for the RAINS model analysis.

The structure of one cycle of the RAINS model analysis is outlined in *Figure 3.3*. Prior to analysis, a data file is prepared that contains all parameters of the RAINS core model. Another data file with a definition of the parameters for a particular scenario is prepared by specialized software. These two data sets are converted by another specialized program into the HDF file. Additionally, a user has the possibility of selecting various options and specifying the corresponding parameters (e.g., of the composite goal function discussed below) and options (e.g., allowing for soft constraints, requesting the balancing of violations with surpluses) that overwrite the default selections and are used for a definition of a particular instance of the non-linear optimization problem.

The optimization problem is generated and solved by the problem-specific model generator linked with a selected non-linear solver library. The generator (the functions of which are described in Section 3.3.2) creates the necessary data structures, which are kept in-core and are used for functions that are required by each of the used solvers. Such an approach allows for the efficient generation and solution of the corresponding large non-linear optimization problem. After an optimal solution is found, a postprocessor converts the solution to a form that is convenient for analysis (by "undoing" the actions of the preprocessor and by computing values of variables, which were not generated).

A solution provided by the postprocessor is processed by a specialized report-writer program, which provides various types of information needed for the analysis of a solution. Afterwards, another scenario is designed based on this analysis and on requests from users. This scenario is used as an input to a new cycle of the analysis.

A particular scenario is defined by many parameters. A minimization of costs related to measures needed for improvement of air quality is a main goal; however, other objectives – such as robustness of a solution, trade-offs between costs and violations of environmental standards – are also important. Therefore, a MCMA has been applied to this case study. A composite goal function (3.4, below) is applied to support the analysis of trade-offs between the following three criteria:

- the minimization of total costs of emissions reduction,
- the minimization of violations of the environmental standards,
- the robustness of solutions.

The following composite criterion function is used:

$$goal\_function = cost + \Theta + \Psi \tag{3.4}$$

where the $cost$ term corresponds to the sum of the costs of emission reductions and Θ and Ψ are regularizing and the penalty term, respectively.

The regularizing term Θ is defined by:

$$\Theta = \epsilon \|z - \bar{z}\|^2 \tag{3.5}$$

where ϵ is a given positive (not necessarily small) number, z denotes a vector composed of decision variables that correspond to emissions, and $\bar{z}$ is a given vector composed of desired (reference) values of emissions. The role of the term Θ is twofold. First, it helps to avoid large variations of solutions (with almost the same value as the original goal function) for problems that differ very little. Second, it substantially improves the numerical stability of the optimization problem. Additionally, the term Θ can be used for the technique called softly constrained inverse simulation. Thus, it is possible to analyze trade-offs between minimization of costs and solutions that correspond closely to various given patterns of emissions defined by $\bar{z}$.

The role of the term Ψ is also twofold. First, it serves as a penalty term for optional variables y, ya, and ye. Second, it provides regularization for these decision variables, which are not covered by the Θ term. The term Ψ is defined by:

$$\Psi = \sum_{l \in L} \sum_{j \in J} \psi(y_{lj}, \rho_o, \sigma_o) + \sum_{j \in J} \psi(ya_j, \rho_a, \sigma_a) + \sum_{j \in J} \psi(ye_j, \rho_e, \sigma_e) \tag{3.6}$$

where $\rho_o, \rho_a, \rho_e, \sigma_o, \sigma_a, \sigma_e$ are given positive parameters, and the function $\psi(\cdot)$ is defined by:

$$\psi(x, \rho, \sigma) = \begin{cases} -\rho\sigma x - \rho\sigma^2/2 & \text{for } x < -\sigma \\ \rho/2x^2 & \text{for } |x| \le \sigma \\ \rho\sigma x - \rho\sigma^2/2 & \text{for } x > \sigma \end{cases} \tag{3.7}$$

Note that $\psi(x, \rho, \sigma)$ is a smooth function that, depending on the parameters ρ and σ, can be used for both purposes that correspond to the role of the term Ψ outlined above. First, it plays the role of a classical quadratic penalty function, if large values of the parameters ρ, σ are selected. Such a function can be used to examine the trade-offs between violations of air quality standards and minimization of costs. Second, it may not be desirable to apply any penalty function for some scenarios in which the balances between violations of environmental targets and surpluses are required. However, in such cases, it is still necessary to apply regularization in order to deal correctly with the soft constraints optionally defined by introduction of decision variables y_{lj}, ya_j, ye_j. A quadratic function is not suitable for this purpose because often violations and surpluses take small values in some grids and large values in other grids; therefore, it is not possible to find a value of the parameter ρ such that it would allow for large values of violations/surpluses in some grids, while serving as a regularizing term for grids where violations/surpluses may be three orders of magnitude smaller, and a specification of different values of ρ for each of about 600 grids is not practicable. Therefore, when used for regularization purposes alone, the function ψ is defined using small values of both parameters ρ, σ, which implies using a flat PWL function with a small quadratic segment needed to make such a function smooth. Finally, the term Ψ provides a similar functionality as the approach commonly known as *soft constraints*.

To summarize the discussion on the form of the goal function (3.4), it is important to stress the fact that a properly defined goal function is the key element for achieving two objectives: namely, (1) providing a tool for a comprehensive problem analysis and (2) assuring possibly good numerical properties of the corresponding optimization problems. The specific form of this function – in particular, the penalty terms for soft constraint violations, the regularizing terms – makes the model analysis very similar to a multi-objective formulation, as applied, e.g., to softly constrained inverse scenario analysis. See Wierzbicki *et al*. (2000) for more details. In the near future the MCMA software, outlined in the next section, will be optionally used for multicriteria analysis as an alternative to the composite goal function.

3.4 Multicriteria Model Analysis

Many papers and books pointed out quite long ago the limitations of the traditional operations research (OR) approach. See, for example, Wierzbicki (1977), Ackoff (1979), Lewandowski and Wierzbicki (1989), Chapman (1992), and Wessels and Wierzbicki (1993). This chapter will briefly summarize only one of these techniques, which seems to be the most natural method that best corresponds to a real-life decision-making process. This is the Aspiration Reservation Based Decision Support (ARBDS) method which is an extension of the *aspiration level* (sometimes referred to as *reference point*) approach, originally proposed by Wierzbicki (1977), later developed and applied in many applications in various fields, and recently presented in detail in Wierzbicki *et al*. (2000). First, however, it is worth pointing out another successful approach, which is based on another constructive proposal for overcoming limitations of the traditional, optimization-centered OR approach. It was formulated by Sawaragi and coworkers in the introduction to their standard textbook on model-based decision support (Sawaragi *et al*., 1985). These concepts have been elaborated on over the years, and one of the streams of the follow-up research is known as the *Shinayakana systems approach* (see, e.g., Sawaragi and

Nakamori, 1991). A more recent overview of this methodology for analyzing complex systems and environmental modeling is presented by Nakamori and Sawaragi (2000).

The following topics are discussed in ensuing subsections:

- basic concepts of MCMA,
- the Aspiration-Reservation method, and
- several basic problems with using weights to convert a multicriteria optimization problem into a parametric single-criterion mathematical programming problem.

3.4.1 Basic concepts of multicriteria model analysis

A key issue in MCMA is the identification and analysis of those parts of the Pareto-optimal solution set that are interesting for the user. Generation and analysis of the entire Pareto set is practically impossible. Therefore, any practical multi-objective method facilitates an analysis of a subset of Pareto solutions that best correspond to preferences of a user.

Multicriteria optimization methods typically assume that a multicriteria problem is converted into an auxiliary parametric single-objective problem whose solution provides a Pareto-optimal point having desired properties. Different methods apply different conversions, but the most commonly known methods can be interpreted (see Makowski, 1994b) in terms of the Achievement Scalarizing Function (ASF). The concept of ASF was introduced by Wierzbicki and it is very useful for comparing different approaches to multicriteria optimization. (See, e.g., Wierzbicki, 1977; Wierzbicki, 1986; and Wierzbicki *et al*., 2000 for mathematical foundations, interpretations, and applications.)

A solution is called a Pareto-optimal (or an efficient) solution, if there is no other solution for which at least one criterion has a better value while values of remaining criteria are the same or better. In other words, one can not improve any criterion without seeing a value of at least one other criterion deteriorate. We refer to properly Pareto-optimal solutions with a prior bound on trade-off coefficients as Pareto solutions (unless otherwise mentioned). A Pareto-optimal point in objective space is composed of values of all criteria for a corresponding Pareto-optimal solution.

The basic distinction between various multicriteria methods is determined by the way in which a method supports selection of Pareto solutions that best correspond to the user's preferences. The aspiration-based methods use the aspiration values – a term used interchangeably with reference point – which are composed of values that a user wants to achieve for each criterion. If a specified aspiration level $\bar{q}$ is not attainable, then the Pareto-optimal point is that nearest to the aspiration level. If the aspiration level is attainable, then the Pareto-optimal point is uniformly better than $\bar{q}$. Properties of the Pareto-optimal point depend on the localization of the reference point (composed of aspiration values) associated with the criteria, and the applied definition of a distance.

To treat attainable aspiration points properly as well, instead of a distance one uses ASF. Therefore the selection of the Pareto-optimal solution depends on the definition of the ASF, which includes a selected aspiration point. Most of those methods use the maximization of an ASF in the form:

$$\mathcal{S}(q, \bar{q}, w) = \min_{1 \leq i \leq n} \{w_i(q_i - \bar{q}_i)\} + \epsilon \sum_{i=1}^{n} w_i(q_i - \bar{q}_i) \tag{3.8}$$

where $q(x) \in R^n$ is a vector of criteria, $x \in X_0$ are variables defined by the core model, X_0 is a set of feasible solutions implicitly defined by the core model, $\bar{q} \in R^n$ is an aspiration point, $w_i > 0$ are scaling coefficients and ϵ is a given small positive number. Maximization of (3.8) for $x \in X_0$ generates a properly efficient solution with the trade-off coefficients (as recomputed in terms of u_i defined below) smaller than $(1 + 1/\epsilon)$. For a non-attainable $\bar{q}$, the resulting Pareto-optimal solution is the nearest (in the sense of a Chebyshev weighted norm) to the specified aspiration level $\bar{q}$. If $\bar{q}$ is attainable, then the Pareto-optimal solution is uniformly better. Setting a value of ϵ is itself a trade-off between getting a too restricted set of properly Pareto-optimal solutions or a too wide set practically equivalent to weakly Pareto-optimal solutions. Assuming the ϵ parameter to be of a technical nature, the selection of efficient solutions is controlled by the two vector parameters: $\bar{q}$ and w.

It is commonly agreed that the aspiration point is a very good controlling parameter for examining a Pareto set (i.e., a set composed of Pareto-optimal solutions). Much less attention is given to the problem of defining the scaling[8] coefficients w. A detailed discussion on scaling coefficients in a scalarizing function is beyond the scope of this chapter. The four commonly used approaches are summarized in Makowski (1994b).

The aspiration-based approaches correspond very well to the concept of satisficing behavior (also called bounded rationality), in which the DM attempts to attain aspiration levels, usually by first trying to improve the criterion that shows the worst performance. (See, e.g., March and Simon, 1958; Wierzbicki, 1982.) This method has several advantages over other multi-objective optimization methods, as discussed in detail in Wierzbicki *et al*. (2000) and is one of the most successful

[8]Note that the scaling coefficients w should not be used as weights for a conversion of a multicriteria problem into a single criterion problem with a weighted sum of criteria (see Section 3.4.3 for a detailed discussion and examples). In the function (3.8) they play a different role than in a weighted sum of criteria.

classes of DSSs (see, e.g., Korhonen and Wallenius, 1989 for a justification of this statement). Moreover, as shown by Ogryczak and Lahoda (1992), the aspiration level approaches may be interpreted as extensions of goal programming (Charnes and Cooper, 1967), which was the precursor of most multicriteria optimization and model analysis methods.

The reference point approach can also be used for inverse simulation: instead of repeatedly adjusting the decision variables in order to determine acceptable states (expressed as constraints in the classical approach to optimization), the user chooses desired states (in terms of ranges of values of objectives) and the DSS determines for her/him the resulting values of the decision variables. The reference point approach also takes into account soft constraints often needed in the single-criterion optimization. Specifically, one can replace a soft constraint (or group of constraints) by an objective, and then set the aspiration level equal to the desired value of the constraint and the reservation level to the worst acceptable value. Thus, violations of soft constraints can be treated as criteria (to be minimized) in the multi-objective approach.

3.4.2 Aspiration reservation based decision support

The ARBDS method is an extension of the reference point method. In practical applications, the most promising approach is based on the calculation of scaling coefficients (used to define the weighted Chebyshev norm mentioned above), with the help of the aspiration level $\bar{q}$ and a reservation level $\underline{q}$ (the latter is composed of values of criteria that the user wants to avoid). This is the ARBDS approach that has been introduced by the DIDAS family of DSS described in Lewandowski and Wierzbicki (1989). Its extension is presented here, applied in the Interactive Specification and Analysis of Aspiration-based user Preferences (ISAAP) tool, which is described in detail in Granat and Makowski (2000).

Geometrical aspects of the reference point and the ARBDS approaches are shown in *Figure 3.4*, which illustrates a Pareto-solution set (between points **D** and **E**) for a two-criteria minimization problem. Utopia and Nadir[9] points (denoted by **U** and **N**, respectively) are composed of the best and worst values of criteria in the Pareto set. For a given aspiration point **A**, any of the Pareto-optimal points between points **B** and **C** can be obtained for various definitions of the distance between an aspiration point and the Pareto set. In the classical reference point method the weights in the ASF defined by (3.8) had to be somehow specified. It

[9]One should note that a computation of a true Nadir point is often practically impossible, therefore, modern MCMA approaches don't use the Nadir point for model analysis. Usually only an approximation of the Nadir point is used for actual model analysis, which is not any actual limitation because users are typically interested in aspiration levels which are not attainable, and extremely bad solutions (which are close to the Nadir point) are not really interesting.

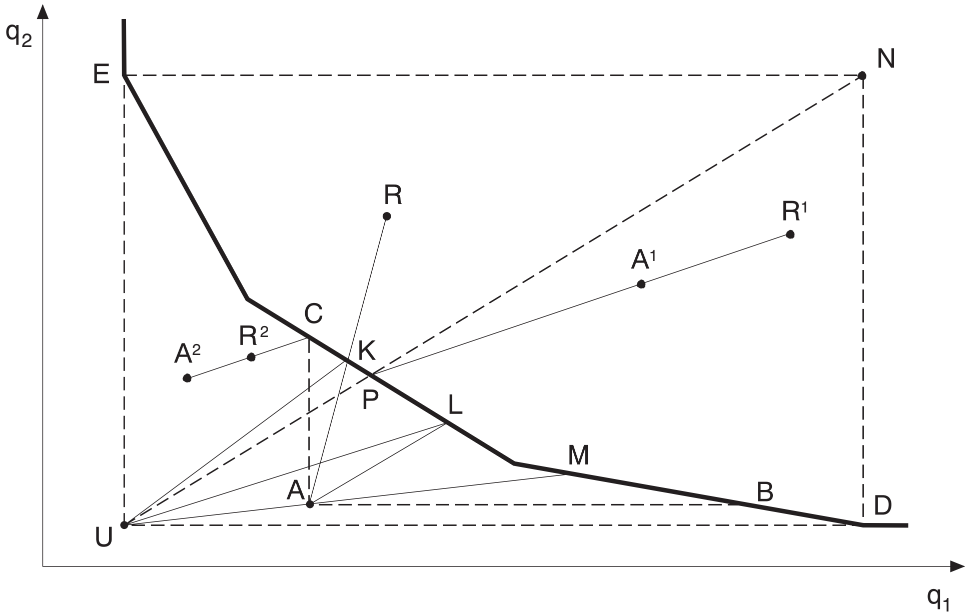


Figure 3.4. Illustration of a Pareto-optimal surface for a two-criteria case.

had been recognized that users have serious difficulties with defining weights that correspond well to their trade-offs between criteria (the problems were similar to those described in Section 3.4.3), and therefore the next stage in the development of the MCMA methods was to define weights indirectly using two points. Two approaches were most popular for a selection of such points: to use either a pair {**U**, **A**}, or {**U**, **N**} for defining the weights. For the example in *Figure 3.4* this results in the Pareto-solutions **M**, and **L**, respectively. However, in practical applications both of these approaches created various difficulties, both for users and for the developers of the needed software.

These difficulties led to the development of the ARBDS method, in which a user specifies for each criterion a pair of values: aspiration (the desired value of a criterion) and reservation (the worst acceptable value), denoted further on by **A** and **R**, respectively. This is the most natural way of examination of interesting parts of the Pareto set because users typically have a good understanding of the range of criteria values that they want to achieve. Moreover, an appropriate implementation of the ARBDS method does not impose any restrictions on **A** nor on **R**. In *Figure 3.4* there are three pairs of aspiration and reservation points, denoted by {**A**, **R**}, {$\mathbf{A}^1$, $\mathbf{R}^1$}, and {$\mathbf{A}^2$, $\mathbf{R}^2$}, respectively. The corresponding Pareto-solutions are marked by **K**, **C**, and **P**, respectively. A selection of a pair like {**A**, **R**} (i.e., a not attainable aspiration and a feasible reservation level) is typical for users who have learned the properties of the problem and have a good feeling about the attainable ranges of criteria values. Selection of non-attainable reservation level (e.g., $\{\mathbf{A}^1, \mathbf{R}^1\}$) is typical for early stages of model analysis, when unrealistic reservation levels are specified. However, specifications of attainable aspiration levels (or at least some components of it) are not as rare as one can expect; especially, if some criteria are correlated. Therefore it is important that MCMA does not impose any restrictions on the feasibility of the aspiration nor of the reservation values.

In order to meet this requirement, and to support an option of a more exact specification of preferences for criteria values between aspiration and reservation values, the ASF for the ARBDS has a more general form than that shown in (3.8, above), and usually is defined as

$$\mathcal{S}(q, \bar{q}, \underline{q}) = \min_{1 \le i \le n} u_i(q_i, \bar{q}_i, \underline{q}_i) + \epsilon \sum_{i=1}^{n} u_i(q_i, \bar{q}_i, \underline{q}_i) \tag{3.9}$$

where $\bar{q}, \underline{q}$ are vectors (composed of $\bar{q}_i, \underline{q_i}$, respectively) of aspiration and reservation levels, respectively, and $u_i(q_i, \bar{q}_i, \underline{q}_i)$ are the corresponding Component Achievement Functions (CAFs), which can be simply interpreted as nonlinear monotone transformations of the i-th criterion value q_i into ASF u_i (which reflects the degree of satisfaction of the user) based on the information represented by aspiration and reservation levels for this criterion, $\bar{q}_i$ and $\underline{q}_i$, respectively. Maximization of the function (3.9) over the set of feasible solutions X_0 defined by the corresponding core model provides a properly Pareto-optimal solution with the properties discussed above for the function (3.8).

The ARBDS method is organized into the following steps typical for a MCMA:

1. The user or DM selects several criteria (objectives) from outcome variables. In typical applications there are 2–9 criteria.
2. The DM specifies an aspiration point $\bar{\boldsymbol{q}} = \{\bar{q}_1, \ldots, \bar{q}_k\}$, where $\bar{q}_i$ are aspiration levels (the desired values for each criterion) and k is the number of criteria. Additionally, the DM specifies a reservation point $\underline{\boldsymbol{q}}$, which is composed of the worst values of criteria that a DM would like to accept. Optionally, the user can specify his/her preferences for values of criteria between aspiration and reservation, by interactively selecting points that define PWL function, as illustrated in *Figure 3.5*.
3. The underlying formulation of the problem is the maximization of a ASF (3.9). This can be interpreted either as a value function of the DSS specified in response to the specific aspiration and reservation levels, or as an ad-hoc, nonstationary approximation of the value function of the DM, dependent on these

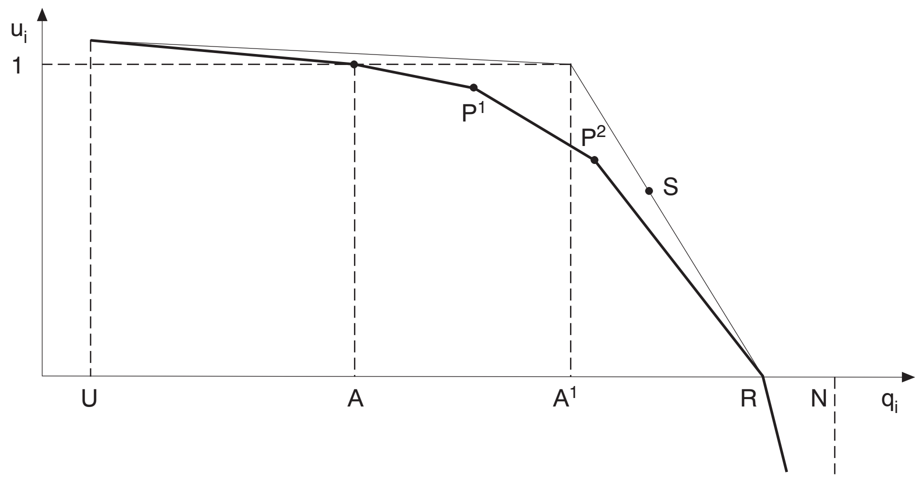


Figure 3.5. Illustration of the PWL CAF for a minimized criterion.

levels. The problem is then transformed by the DSS into an auxiliary parametric single-objective problem, the solution of which gives a Pareto-optimal point.

4. The DM explores various Pareto-optimal points by changing the aspiration point $\bar{q}$ and reservation point $\underline{q}$ for each criterion. Additionally, a DM may stabilize a criterion (i.e., specify a desired value instead of minimizing or maximizing the value of this criterion) or temporarily remove a criterion from the analysis.
5. The procedure described in points 2, 3, and 4 is repeated until a set of satisfactory solutions is found.

In order to allow for either specification of only aspiration and reservation levels or for additional specification of preferences (for the criteria values between aspiration and reservation levels), the ISAAP supports specification of the CAFs in a more general form than that originally proposed by Wierzbicki (1986). For this purpose, the PWL CAF u_i are defined by segments u_{ji}:

$$u_{ji} = \alpha_{ji} q_i + \beta_{ji}, \quad q_{ji} \leq q_i \leq q_{j+1,i} \quad j = 1, \ldots, p_i \quad p_i \geq 3 \tag{3.10}$$

where p_i is a number of segments for the i-th criterion. Such a function for a minimized criterion is illustrated in *Figure 3.5*. The thin line corresponds to a function that is composed of three segments, which are defined by four points, namely **U**, $\mathbf{A}^1$, **R**, and **N** (corresponding to the Utopia, aspiration, reservation, and Nadir points, respectively). The solid line represents a modified function for which the previously defined aspiration level $\mathbf{A}^1$ was moved to the point **A** and two more points – $\mathbf{P}^1$ and $\mathbf{P}^2$ – were interactively defined.

Values of CAF have a very easy and intuitive interpretation in terms of the degree of satisfaction from the corresponding value of the criterion. Values of 1 and 0 indicate that the value of the criterion exactly meets the aspiration and reservation values, respectively. Values of CAF between 0 and 1 can be interpreted as the degree of *goodness* of the criterion value, i.e., to what extent this value is close to the aspiration level and far away from the reservation level. These interpretations correspond to the interpretation of the membership function (MF) of the Fuzzy Sets, which is discussed in the section on the relations between fuzzy sets and CAF, below. However, the CAF values provide more functionality than the MF of the Fuzzy Sets, which does not distinguish the differences between elements that do not belong to a set. Namely, values of CAF greater than 1 correspond to the criterion values better than aspiration level while negative values of CAF show that the criterion value is worse than the reservation level, and the differences in values of CAF correspond to the differences of quality of solutions that are beyond aspiration and reservation levels.

By using an interactive tool for specification of the CAF defined by (3.10), such as ISAAP (Granat and Makowski, 2000), a DM can analyze various parts of a Pareto set that best correspond to various preferences of trade-offs between criteria. These preferences are typically different for various stages of analysis, and are often modified substantially during the learning process, when aspiration and reservation levels for criteria values are confronted with the attainable solutions, which correspond best to the aspiration and reservation levels. In such an interactive learning process, a user gradually comes to recognize attainable goals that correspond best to his/her trade-offs.

In some applications, the value of an outcome variable should neither be minimized nor maximized but should have a value close to a given *target* value. In such a situation, the goal-type criterion can be used. For this type of a criterion, the distance from a given target value (which can be changed during the interaction) is to be minimized. For such a criterion, a CAF is composed of two parts: the first part is defined for the criterion values smaller than the target value, and the second part for the criterion values larger than the given target. Such a function is illustrated in *Figure 3.6*. The conditions specified above for maximized and minimized criteria hold for the first and second function, respectively. There is obviously only one point i, for which $\alpha_{i-1,i} > 0$ and $\alpha_{i,i+1} < 0$ and the criterion value for such a point corresponds to a target value (denoted by **T** in *Figure 3.6*) for the goal-type criterion. The function shown in *Figure 3.6* is symmetric, but for many applications an asymmetric function is appropriate and therefore both types of functions

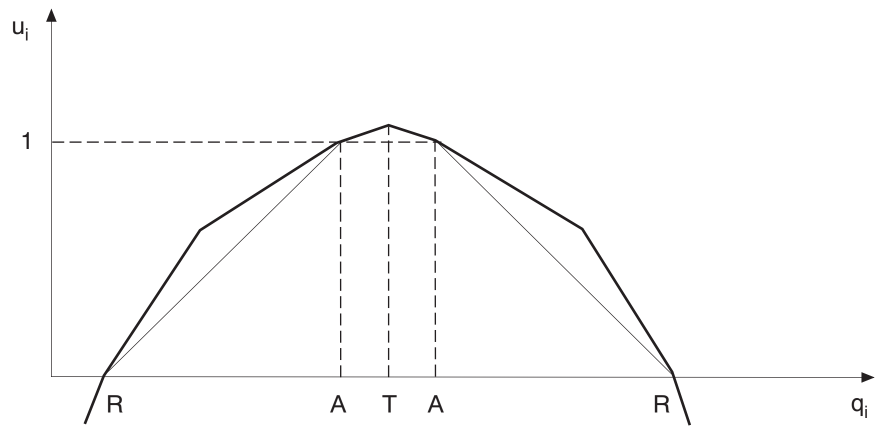


Figure 3.6. Illustration of the PWL goal-type CAF for a goal-type (or stabilized) criterion.

for the goal-type of criteria are supported by ISAAP. More details, including also a discussion on asymmetric CAF can be found in Granat and Makowski (1998).

The three types of criteria, i.e., minimized, maximized and goal-type, are used most often. However, sometimes it is desirable to consider more complex criteria. For example, for dynamic problems it is typical to deal with trajectories. In such cases, one can easily define outcome variables that correspond to a deviation from a given trajectory. Depending on the application, either a trajectory should be followed, or only the surplus (or deficit) should be minimized. The corresponding outcome variables can be defined as follows:

$$\max_{t \in T} \mid x_t - \bar{x}_t \mid \tag{3.11}$$

$$\max_{t \in T}(x_t - \bar{x}_t) \tag{3.12}$$

$$\min_{t \in T}(x_t - \bar{x}_t) \tag{3.13}$$

where T is a set of time indices and $\bar{x}_t$ is a given reference trajectory. Such variables can be used as criteria: minimized for the first two cases and maximized for the last case.

The maximization of the ASF defined by (3.9) which uses the CAF in the form of (3.10) provides a natural way for selecting Pareto-efficient solutions that conform to the concept of satisficing behavior, that is, situations in which the DM attempts to attain aspiration levels by first trying to improve the criterion that shows the worst performance, i.e., which differs most from its aspiration level.

Relations Between Fuzzy Sets and Component Achievement Function

This section briefly comments upon an interpretation of the ASF in terms of Fuzzy Sets. Such functions can be interpreted in terms of the fuzzy MFs discussed in detail in Zimmermann (1985). MF are typically interpreted as functions that reflect the degree to which an element belongs to a set.

The ARBDS approach uses a so-called extended-valued MF proposed by Granat and Wierzbicki (1994), who suggested a method for constructing various forms of order-consistent component ASFs based on MFs that describe the satisfaction of the user with the attainment of separate objectives. Between aspiration and reservation levels, the values of this function coincide with the MF, as well as having an ordering properties. In other segments an ASF is used only for ordering alternatives (thus assuring that only Pareto-efficient solutions are found).

Thus, there are many similarities between the ARBDS and the *Fuzzy Multi-objective Programming* approaches. The main difference is due to the specification and use of CAF. The Fuzzy Multi-objective Programming method requires prior specification of aspiration and reservation levels that are used to define the MFs. It is implicitly assumed that the criteria values for all the interesting solutions are between the corresponding aspiration and reservation levels (because the applied MF does not differentiate between solutions with values better than aspiration level and between those with values worse than reservation level). In MCMA the user interactively specifies the reference membership levels for each CAF, which can be interpreted as a degree of achievements of the aspiration for each criterion. Moreover, CAF has order-preserving property, i.e., it has different values for all different criterion values.

The ARBDS method does not use the MF directly. It assumes that the user may change aspiration and reservation levels during the interaction upon the analysis of previously obtained solutions. The user specifies interactively the preferences in the space of the criteria values, which seems to be more natural than a specification of preferences in terms of degrees of achievements of CAF values. A selection in the criteria space can, however, be interpreted in terms of Fuzzy Sets by a definition of an MF for a linguistic variable (e.g., *good solution*) for each criterion, and an ex-post interpretation to which degree a solution belongs to a set of *good* solutions. There is no need for restrictions for the specification of aspiration and reservation levels in the criteria space. This is important for the analysis of large-scale complex problems for which the specification of attainable reservation levels might be difficult.

3.4.3 Problems with using weights for aggregation of criteria

One of the most popular approaches to multicriteria optimization is based on the idea of converting a multicriteria problem into a single-criterion one by summing up weighted criteria. This approach has a number of drawbacks as discussed in detail, e.g., by Makowski (1994b), by Nakayama (1994) and in Wierzbicki *et al*. (2000). However, this approach is still popular because it is believed to be simple, intuitive, and reliable. Thus it is necessary to summarize here the main limitations and misinterpretations of the properties of this approach.

First, there are fundamental problems with determining correct weights. One can easily observe via a simple example that weights (which always attempt to reflect a relative importance of criteria) must be in this approach defined as parameters that are constant for the whole Pareto set. However, the weights are actually very different in various areas of a Pareto set. To illustrate this point let us consider two minimized criteria:

- q_1, costs of emission reduction, and
- q_2, a measure of a concentration of pollution,

and the corresponding weights w_1 and w_2. For two-criteria examples it is enough to consider the ratio:

$$\alpha = w_1/w_2. \tag{3.14}$$

Typically, when q_1 attains its best value (which corresponds to a minimum cost solution, which results in a high value of q_2) the value of α will be rather low, indicating much higher weight attached to the environmental criterion. An application of a very small α for the example presented in *Figure 3.7a* would result in the solution **A**. Conversely, for a best available purification technology the q_2 will attain minimum, which also corresponds to highest costs. In such a situation α will take a rather high value corresponding to a much higher weight attached to the economic criterion and the selected solution will be at point **C**.

Second, a weighted aggregation of criteria is a very unreliable way of scanning a set of Pareto solutions. Consider the simplest case with two minimized objectives illustrated in *Figure 3.7a*. For the linear case, a user can obtain only Pareto-optimal solutions corresponding to vertices **A**, **B**, and **C**. For any weighting coefficients vector α with a slope smaller than the slope of the vector α^1, a solution will be in the vertex **A**. For a weighting coefficient vector that is parallel to α^1, there is no unique solution,[10] and a very small increase of the slope of α will cause the solution to

[10]Therefore the corresponding problem will be degenerated and any solution from the edge **AB** is optimal. Hence, the reported solution will differ, depending not only on the applied solver but also on the parameters used for a solver, including the possibly defined starting point for optimization.

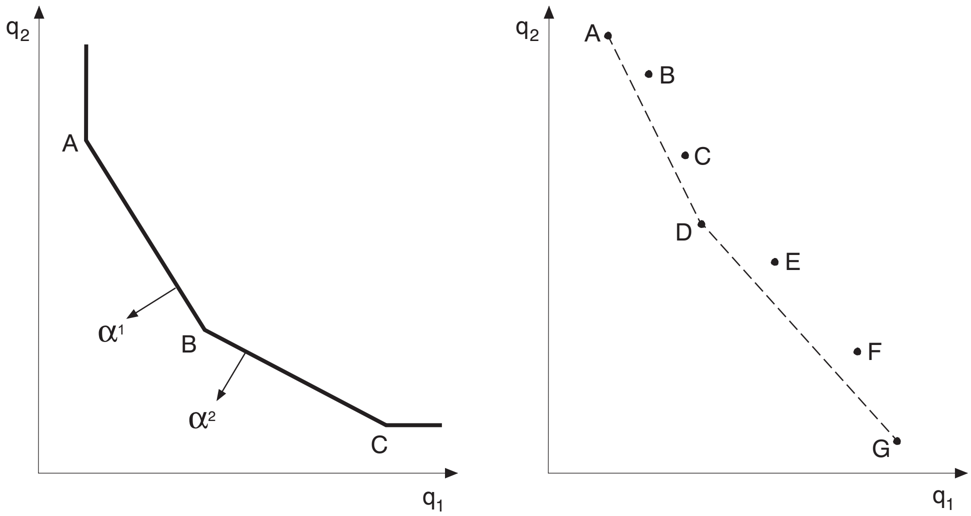


Figure 3.7. Limitations of selecting all Pareto solutions by aggregating criteria through their weighted sum: the cases of continuous (a) and discrete (b) linear models.

jump to the vertex **B**. A further increase in the slope of α will not cause any changes in the Pareto solution until the slope becomes greater than α^2 (which will cause another jump to the vertex **C**). This explains the experience known to everyone who has tried to use weights to analyze multiple-criteria LP models, namely, that often a relatively large change of weights does not cause any changes to the solution, but for some combinations of weights, a small modification creates in the same model a substantially (in practice the distances between vertices are often large) different solution.

Third, a weighted aggregation of criteria does not allow us to find all Pareto solutions. For a discrete model, a surface spanned over the Pareto set (that is composed of points) may be non-convex. Therefore, for the example depicted in *Figure 3.7b*, only some efficient solutions, namely, **A**, **D**, **G** will be found while possibly many other efficient solutions (e.g., **B**, **C**, **E**, **F**) will never be found, if weights are applied for an aggregation of criteria.

Fourth, contrary to the common belief, using weights can be counterintuitive, as one can find examples in which, for certain regions of the efficient frontier, there is no positive correlation between increasing the weight for a criterion and the corresponding improvement of the criterion value. Nakayama (1994) provides a more formal discussion of this issue illustrated by an example that shows that there might be no positive correlation between increasing a weight for a criterion and the corresponding improvement of the criterion value.

Other problems and limitations of using weights for aggregating criteria are discussed by K. Hayashi and by A. Mohamed, in Chapters 4 and 13 of this volume, respectively.

Given such serious problems with using weights, there is no justification to use this approach any longer, especially because more reliable and natural approaches to MCMA are easily available.

3.5 Conclusions

Modeling of complex systems does, and will, require various elements of science, craftsmanship, and art (see, e.g., Wierzbicki *et al*., 2000 for a collection of arguments that supports this statement). Moreover, development and comprehensive analysis of a complex model requires – and will continue to require – a substantial amount of time and resources. The main message of this chapter is to stress the often-forgotten fact that no single modeling paradigm can be successfully used to analyze a complex problem, especially if the results of such an analysis are used to support various elements of real decision-making processes. Several rules must be observed during the specification of a model in order to provide useful results. Also, various techniques of model analysis should be used, rather than just the classical approaches which are focused and driven either by simulation or optimization paradigms. Finally, the use of modular re-usable software tools should also be emphasized. They substantially ease the implementation of DSSs that provide more complete and comprehensive analysis of a problem than do closed systems focused on a specific model-analysis paradigm.

References

Ackoff, R., 1979, The future of operational research is past, *Journal of OR Society*, **30**(2):93–104.

Alcamo, J., Shaw, R., and Hordijk, L., eds, 1990, *The RAINS Model of Acidification*, Kluwer Academic Publishers, Dordrecht, Boston, London.

Amann, A., and Makowski, M., 2000, Effect-focused air quality management, in A. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 367–398. ISBN 0-7923-6327-2.

Chapman, C., 1992, My two cents worth on how OR should develop, *Journal of Operational Research Society*, **43**(7):647–664.

Charnes, A., and Cooper, W., 1967, *Management Models and Industrial Applications of Linear Programming*, J. Wiley & Sons, New York, London.

Drud, A., 1997, Interactions between nonlinear programming and modeling systems, *Mathematical Programming*, **79**(1-3):99–123.

Goodwin, P., and Wright, G., 1991, *Decision Analysis for Management Judgment*, J. Wiley & Sons, Chichester, New York.

Granat, J., and Makowski, M., 1998, ISAAP – Interactive Specification and Analysis of Aspiration-Based Preferences, *Interim Report* IR-98-052, International Institute for Applied Systems Analysis, Laxenburg, Austria. Available on-line from http://www.iiasa.ac.at/∼marek/pubs

Granat, J., and Makowski, M., 2000, Interactive Specification and Analysis of Aspiration-Based Preferences, *European Journal of Operations Research*, **122**(2):469–485. Available also from International Institute for Applied Systems Analysis, Laxenburg, Austria, as RR-00-09.

Granat, J., and Wierzbicki, A. P., 1994, Interactive specification of DSS user preferences in terms of fuzzy sets, *Working Paper WP-94-29*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Granat, J., Makowski, M., and Wierzbicki, A., 2000, Optimization tools, in A. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 167–214. ISBN 0-7923-6327-2.

Korhonen, P., and Wallenius, J., 1989, Observations regarding choice behaviour in interactive multiple criteria decision-making environments: An experimental investigation, *in* A. Lewandowski and I. Stanchev, eds, *Methodology and Software for Interactive Decision Support*, Vol. 337 of *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin, New York.

Koziol, Q., and Matzke, R., 1998, *HDF5 – A New Generation of HDF: Reference Manual and User Guide*, National Center for Supercomputing Applications, Champaign, Illinois, USA, http://hdf.ncsa.uiuc.edu/nra/HDF5/.

Lewandowski, A., and Wierzbicki, A., eds, 1989, *Aspiration Based Decision Support Systems: Theory, Software and Applications*, Vol. 331 of *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin, New York.

Makowski, M., 1994a, Design and implementation of model-based decision support systems, *Working Paper WP-94-86*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Makowski, M., 1994b, Methodology and a modular tool for multiple criteria analysis of LP models, *Working Paper WP-94-102*, International Institute for Applied Systems Analysis, Laxenburg, Austria. Available on-line from http://www.iiasa.ac.at/∼marek/pubs

Makowski, M., 2000, Modeling paradigms applied to the analysis of European air quality, *European Journal of Operations Research*, **122**(2):219–241. Available also as from International Institute for Applied Systems Analysis, Laxenburg, as RR-00-06.

Makowski, M., and Somlyódy, L., 2000, River basin water quality management, in A. Wierzbicki, M. Makowski, and J. Wessels, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 311–332. ISBN 0-7923-6327-2.

March, J., and Simon, H., 1958, *Organizations*, J. Wiley & Sons, New York.

Nakamori, Y., and Sawaragi, Y., 2000, Complex systems analysis and environmental modeling, *European Journal of Operations Research*, **122**(2):178–189.

Nakayama, H., 1994, Aspiration level approach to interactive multi-objective programming and its applications, *Working Paper WP-94-112*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Ogryczak, W., and Lahoda, S., 1992, Aspiration/reservation-based decision support — a step beyond goal programming, *Journal of Multi-Criteria Decision Analysis*, **1**(2):101–117.

Olendrzyński, K., Berge, E., and Bartnicki, J., 2000, EMEP eulerian acid deposition model and its application, *European Journal of Operations Research*, **122**(2):426–439.

Ryoke, M., Nakamori, Y., Heyes, C., Makowski, M., and Schöpp, W., 2000, A simplified ozone model based on fuzzy rules generation, *European Journal of Operations Research*, **122**(2):440–451. Available also from International Institute for Applied Systems Analysis, Laxenburg, Austria, RR-00-07.

Sawaragi, Y., and Nakamori, Y., 1991, An interactive system for modeling and decision support – Shinayakana system approach, *in* M. Makowski and Y. Sawaragi, eds, *Advances in Methodology and Applications of Decision Support Systems*, *Collaborative Paper* CP-91-17, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Sawaragi, Y., Nakayama, H., and Tanino, T., 1985, *Theory of Multiobjective Optimization*, Academic Press, New York.

Schöpp, W., Amann, M., Cofała, J., and Klimont, Z., 1999, Integrated assessment of European air pollution emission control strategies, *Environmental Modelling and Software*, **14**(1):1–9.

Wessels, J., and Wierzbicki, A., eds, 1993, *User-Oriented Methodology and Techniques of Decision Analysis and Support*, Vol. 397 of *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin, New York.

Wierzbicki, A., 1977, Basic properties of scalarizing functionals for multiobjective optimization, *Mathematische Operationsforschung und Statistik, s. Optimization*, **8**:55–60.

Wierzbicki, A., 1982, A mathematical basis for satisficing decision making, *Mathematical Modelling*, **3**(5):391–405.

Wierzbicki, A., 1986, On the completeness and constructiveness of parametric characterizations to vector optimization problems, *OR Spektrum*, **8**:73–87.

Wierzbicki, A., Makowski, M., and Wessels, J., eds, 2000, *Model-Based Decision Support Methodology with Environmental Applications*, Series: Mathematical Modeling and Applications, Kluwer Academic Publishers, Dordrecht. ISBN 0-7923-6327-2.

Wolfram, S., 1996, *The Mathematica Book, Third Edition, Mathematica Version 3*, Cambridge University Press, Cambridge.

Zimmermann, H., 1985, *Fuzzy Set Theory – and Its Applications*, Kluwer Academic Publishers, Boston, Dordrecht, Lancaster.

Chapter 4

On the Applicability of Multicriteria and Risk Analysis to Agri-Environmental Policy Making

Kiyotada Hayashi

Abstract

This chapter reviews methodologies used for analyzing agri-environmental problems, especially for evaluating agricultural practices. Its purpose is to provide systematic perspectives on the problems in which diversified research topics have been discussed. First, multicriteria analysis applied to the problems of selecting farming practices is surveyed and the difficulties in utilizing the methodology, which are related to the weighting procedure in multiattribute value functions, are outlined. Second, the applicability of risk concepts for health and ecological issues is examined in order to resolve the difficulties; a mapping technique is used for clarifying the relationship between agricultural practices and the environment. Problems with valuations of human health and the environment are also presented. Finally, the weighting problem is revisited on the basis of the risk concepts and a dilemma in applying economic evaluation methods and decision analytic approaches is discussed.

Keywords: Agricultural practices, environmental impact, health and ecological risks, multiattribute value functions, risk-benefit analysis.

4.1 Introduction

The impact of agriculture on the environment attracts public attention because agricultural practices involve, in many cases, the degradation of environmental quality. One of the serious problems that can be observed worldwide, for example, (Fried,1991; Heathwaite *et al.*, 1993; Addiscott, 1996; Kumazawa, 1999; Lægreid *et al.*, 1999) is water pollution caused by farming practices, notably the nitrate issue caused by chemical fertilizers and manure, although agriculture is not the only source of the contamination. These concerns necessitate considering economic-environmental tradeoffs, although interactions between humans and the environment are necessary for preserving biodiversity.

Since availability of sufficient data that are area-specific is in most cases essential to studying the possibility of alternative agricultural practices, various studies of farming and cropping systems based on field experiments or simulation have thus been performed while considering their effects on the environment (Kelly *et al*., 1996; Bailey *et al*., 1999). In addition to the difficulty in obtaining sufficient and reliable data, there is wide diversity among the methodologies used for analyzing agri-environmental problems. Developing agri-environmental indicators is one method for identifying and quantifying the extent of the impacts of agricultural policies on the environment (OECD, 1997). Environmental life cycle assessment (LCA) is another method for integrating various impacts on the environment and it has been applied to agricultural products (Sleeswijk *et al*., 1996; Audsley, 1997).

However, since huge research topics as well as various research methodologies are discussed in these fields, it is necessary to provide systematic perspectives. Multicriteria analysis (including utility theory) can be considered as one of the methods that will provide such views. Actually, multicriteria analysis has been studied in the evaluation phase of environmental life cycle assessment (Heijungs, 1992a, 1992b); moreover, multiattribute value functions, for example, have been used for selecting agricultural practices (Hayashi, 2000a). This kind of approach is especially important in considering recommended agricultural practices such as Good Agricultural Practices, Best Agricultural Practices, and Best Management Practices (Bertilsson, 1992; Croll, 1994; California Fertilizer Association, 1995), because investigation of agricultural systems inevitably involves environmental consideration.

Therefore, this chapter rethinks the appropriateness of multicriteria analysis applied to agri-environmental problems and discusses the possibility of applying risk concepts to health and ecological issues in improving the usefulness of the methodology. In Section 4.2, multicriteria analysis used for evaluating agricultural practices is described and difficulties are outlined through a review of previous studies. In Section 4.3, the applicability of health and ecological risks is discussed after restructuring the problem.

4.2 Multicriteria Analysis for Agricultural Practices

4.2.1 Selecting farming practices by multiattribute models

This section concentrates on using multicriteria analysis to select agricultural practices expressed as discrete alternatives, although various multicriteria analyses have been applied to agricultural and natural resource management (Hayashi, 1999a, 2000b). As shown in *Table 4.1*, there are two types of methods in the applications. One is the compensatory approach which aggregates multiple attributes into overall values by, e.g., multiattribute value (utility) functions in which the concept of tradeoffs plays a crucial role. The other is the non-compensatory or outranking approach which introduces aggregation procedures based on concordance and discordance concepts that are derived from outranking relations, which express that an alternative is at least as good as another one. Although the distance-based approach (e.g., compromise programming), in which the distance between the ideal point and the alternatives is minimized, can also be used for decision making, this table contains no examples of that approach.

One of the main features in these applications is that attention is paid to the tradeoffs between economic objectives and environmental objectives [except for Arondel and Girardin (1998)]. That is, most of the problems can be expressed hierarchically as depicted in *Figure 4.1*; agricultural practices are evaluated from the viewpoint of profitability and environmental quality of soil and water.

As a representative method for analyzing the tradeoff, we focus on a method based on an additive multiattribute value function used for ranking discrete alternatives. This method has been applied to many selection problems of farming practices, and has been used in many papers presented at both the First International Conference on Multiple Objective Decision Support Systems (MODSS) for Land, Water, and Environmental Management in 1995 (El-Swaify and Yakowitz, 1998) and the second such conference in 1999 (MODSS‘99, 1999).

This method is based on an importance order of attributes which is elicited from decision makers or experts without specifying numerical values of attribute weights and has the following procedures. First, single-attribute values are calculated using value functions. Next, the priority order of the attributes is given. Then, the best

Table 4.1. Criteria used for selecting farming practices.

Authors	Economic	Fertilizer	Environmental Pesticide	Other
Yakowitz *et al.* (1993)	Net income	N (percolation) N (surface) P (surface)	Atrazine (surface) Atrazine (percolation) Serin (surface) Carbofuran (surface)	Sediment yield
Foltz *et al.* (1995)	Net returns	N (surface) N (percolation) [EPIC]	Atrazine (surface) Alachlor (surface) [GLEAMS]	Soil loss [USLE]
Heilman *et al.* (1997)	Net returns	N (runoff) NO_3-N (percolation)	Atrazine (runoff) Atrazine (sediment) All other pesticides in surface or groundwater	Soil detachment Sediment yield
Lawrence *et al.* (1997)	Above ground net primary production			Range condition Channel erosion Annual runoff Annual maximum Peak runoff rate Quail and javalina (NRCS wildlife habitat index)
Arondel and Girardin (1998)		N management (amount, balance, date, splitting up, improving techniques)	Pesticide management (amount, half-life, mobility, toxicity, location, date)	Water management (hydric balance, amount
Tiwari *et al.* (1999)	Farmers' NPV Government NPV Societal NPV			Land suitability Energy output/input Water requirement Environmental cost

and worst overall values are determined by solving the following linear programs:

$$\text{minimize or maximize} \quad v(a_i) = \sum_{j=1}^{n} w_j v_j(x_{ij})$$

$$\text{subject to} \quad \sum_{j=1}^{n} w_j = 1,$$

$$w_1 \geq w_2 \geq \ldots \geq w_n \geq 0,$$

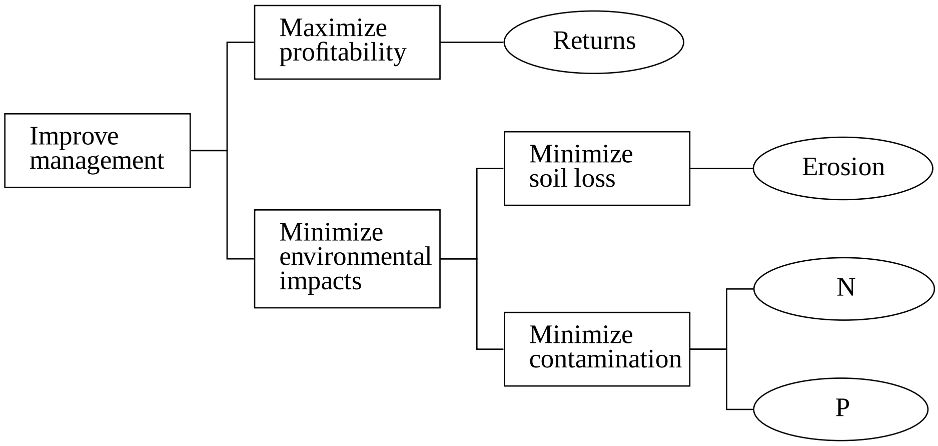


Figure 4.1. An example of a value tree for agricultural practices.

where $v(a_i)$ is the overall value of alternative a_i; w_j is the weight for the jth attribute; $v_j(\cdot)$ is the value function for the jth attribute; x_{ij} is the jth attribute level for alternative a_i; and n is the number of attributes.

This method, however, is associated with the following problems (Hayashi, 1999b). (1) The meaning of rank-ordered weights based on the relative importance of attributes is ambiguous; moreover, the weights based on importance judgments may distort rescaling of single-attribute value functions. (2) The importance order of attributes may not be sufficient to determine the final ranking of alternatives, because the intervals of overall values obtained from the importance order are wide and overlap each other. (3) The fact that the structure of a value tree (a hierarchy of criteria) may affect the final results is not taken into account.

4.2.2 Issues on attribute weights

One of the methods to remedy the first problem, the most serious difficulty, is the use of difference value measurement (Dyer and Sarin, 1979) and the application of weight elicitation techniques based on the measurement. It is well-known that using weight elicitation methods without relying on attribute ranges might lead to biased weights (von Nitzsch and Weber, 1993; Fischer, 1995).

This insight is very important because similar procedures have been employed in many fields. For example, multiattribute evaluation techniques are used to integrate geographical data into spatial decision making using GIS, and the range problem just mentioned has been recognized as a common source of error (Malczewski, 1999). Moreover, weighting is a subjective and crucial part in LCA (Goedkoop *et* *al*., 2000) and thus weighting steps are referred to as an "optional element" in ISO 14042 (2000).

However, the weights for attributes expressed as raw data (e.g., the levels of NO_3-N in groundwater) are sometimes difficult for decision makers – and even for experts – to understand. Moreover, the difficulty in weighting when the problems have 10 attributes or more is pointed out in LCA (Goedkoop *et al*., 2000). Indeed, the problems in the real world tend to have a considerable number of attributes.

Therefore, it is necessary to introduce a methodology for transforming the data into the other values to make the meaning easy to grasp and to reduce the number of attributes using a common physical unit. This is especially true for societal decision making because in order to settle the differences in perceptions, it is necessary to introduce an understandable scale into evaluation processes.

4.3 Reformulating the Problem

4.3.1 Restructuring by a mapping method

As a methodology to consider the relationship between agricultural practices and the environment, a mapping method (Banxia Software, 1999) is used in this section because it can graphically represent the complex relations among actions, phenomena, and concepts.

Figure 4.2 shows an example of a cognitive map. This figure shows that practices such as fertilizer application cause many effects on the environment. This diagram expresses one of the possible overviews constructed on the basis of previous studies (Matson *et al*., 1997; Vitousek *et al*., 1997; Lægreid *et al*., 1999; Tilman, 1999). Although this figure does not show the influence on production costs, it is necessary to consider the influence when evaluating agricultural practices.

The main features of this figure are summarized as follows. (1) The attributes employed in the evaluation of agricultural practices explained in the previous section are recognized as intermediate attributes. Consequently, those attributes may be inappropriate in evaluating agricultural practices, although calculating the risks is in general not easy because of the difficulty in obtaining sufficient data. (2) Many attributes depicted in this figure give rise to the two important concepts: human health risk and environmental or ecological risk. Protection of human health and protection of the environment can be considered as dual goals of laws and regulations that use risk assessment to inform decision making (The Presidential/Congressional Commission on Risk Assessment and Risk Management, 1997b).

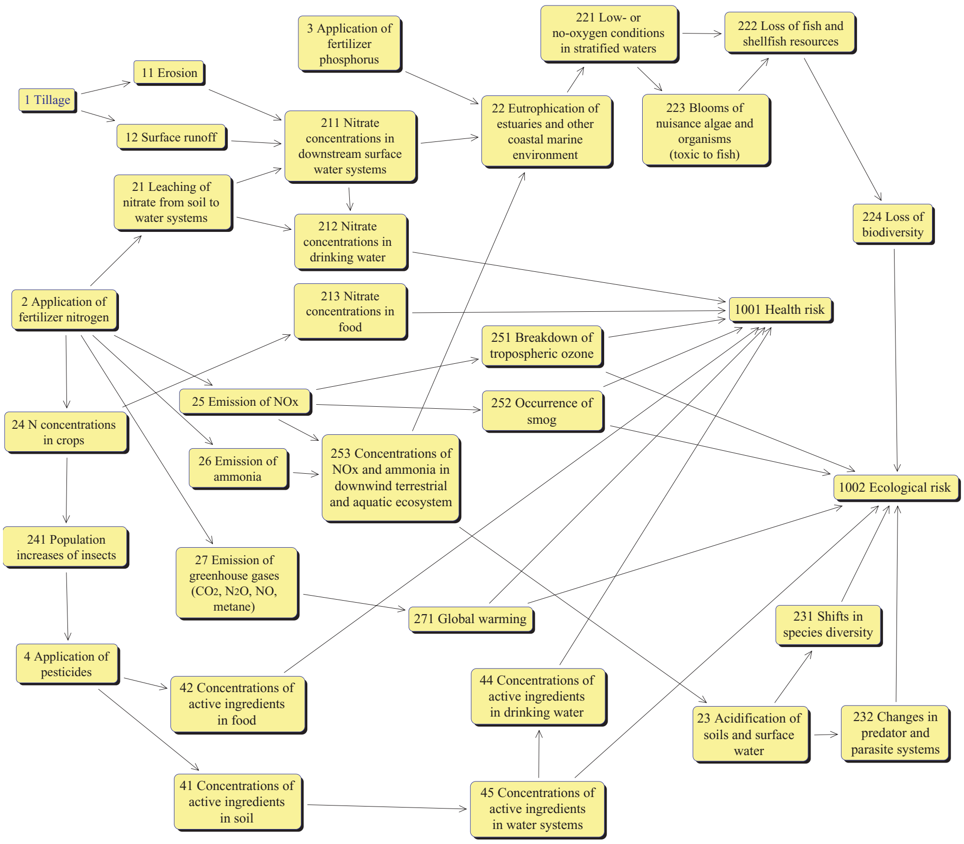


Figure 4.2. A cognitive map.

4.3.2 Health and ecological risks

Having clarified the pervasiveness of the two risk concepts, our next task will be to examine what kinds of units are used for estimating each risk. In the following section, the current situation of these fields is outlined, rather than concentrating on the detailed measurement and calculation processes.

Valuation of human health

A standard economic tool used for evaluating, for example, health care programs is cost-benefit analysis (CBA). In CBA all benefits are measured in terms of money and thus the results obtained from CBA are used to make decisions as to whether to fund a program based on analysis of a value of net benefits (benefits minus costs) expressed in monetary units.

There are, however, difficulties and controversies in assigning a monetary value to human life (a value of a statistical life) or to a change in the quality of life. Concerns have been expressed that economic analysis places too much emphasis on assigning monetary values to aspects of health and the environment that are difficult to quantify in monetary terms (The Presidential/Congressional Commission on Risk Assessment and Risk Management, 1997a). Thus, in health economics, risk analysis, and the life cycle impact assessment, scales constructed from life years have become common practices as follows.

Health economics uses quality-adjusted life years (QALYs) to address changes in length of life and in quality of life (Dasbach and Teutsch, 1996). The number of QALYs is calculated as the sum of the years of life in each health state times the quality of life in each health state. QALYs provide a method for comparing the alternatives (interventions) that have widely varying outcomes; thus they are useful for cost-effectiveness analysis (CEA) and cost-utility analysis (CUA). CUA can be recognized as a special kind of CEA when alternatives have multiple outcomes, although not all researchers accept the term CUA (Gold *et al*., 1996). Although similar concepts, such as the Quality of Well-Being (QWB) and the Health Utility Index (HUI) have been proposed, they are not widely applied.

In addition to QALYs, disability-adjusted life years (DALYs) have also been introduced into the life cycle impact assessment (Hofstetter, 1998; Goedkoop and Spriensma, 2000). DALYs are defined as the sum of the years of life lost (YLL) due to premature mortality and the years lived with disability (YLD) (Murray, 1994; Hofstetter, 1998). This concept is used in international organizations such as the World Health Organization and the World Bank. Moreover, in risk assessment, a loss of life expectancy (LLE) is used to integrate the magnitude of cancer risk and noncancer risk (Nakanishi and Gamo, 1998).

Valuation of the environment

As compared with measurement of benefits from health programs, which is based on the concept of a statistical life (including a risk component) and on survey techniques using willingness-to-pay, valuation of the environment has considerable difficulty. First, the monetary evaluation method has been applied without physical concepts. Indeed, contingent valuation surveys are recognized as notoriously unreliable, especially when applied to issues with which the public is unfamiliar (Daily *et al*., 2000). Ecosystems are, in general, poorly known and are likely to remain elusive.

Moreover, the fact that areas where issues concerning biodiversity are raised are often located next to residential areas makes valuation tasks more complicated. The problem is that the evaluation of biodiversity becomes a secondary concern; the appropriate interaction between humans and nature naturally gains ascendency.

Recent developments in evaluation methodologies are related to the above discussion. The diversity of species is considered to adequately represent the quality of ecosystems in the life cycle assessment (LCA) context (Goedkoop and Spriensma, 2000), as well as in the risk-benefit context (Oka *et al*., 1999). In the former Eco-indicator 99 impact assessment methodology, PAF [the Potentially Affected Fraction of species, which can be interpreted as the fraction of species that is exposed to a concentration equal to or higher than the No Observed Effect Concentration (NOEC)] is used for toxicity, and PDF (the Potentially Disappeared Fraction, which can be interpreted as the fraction of species that has a high probability of no occurrence in a region due to unfavorable conditions) is utilized for acidification, eutrophication, and land-use (Goedkoop and Spriensma, 2000). In the latter analysis, the index of expected loss of biodiversity (ELB) is developed (Oka *et al*., 1999). ELB can be defined as the weighted sum of the increments in the probabilities of extinction of species that would be caused by human activities such as land-use conversion or pollution.

4.3.3 Weighting revisited

As shown in the earlier discussion, the concept of health and ecological risks plays an important role in translating the various consequences. At the current level of risk assessment methodology and scientific knowledge, however, it may be difficult to integrate all the environmental indicators into a physical scale on the basis of the risks. Therefore, we will review the weighting problem once again. First, we discuss attributes like QALYs; then, issues on the integration of attributes are treated.

Concepts like QALYs are based on multiattribute models. This means that from a decision analytic perspective, (1) QALYs can be considered as constructive attributes (Keeney, 1992) and (2) assigning preferences to health states is equivalent to organizational or societal decision making, and thus it is inevitable that we introduce a methodology that is different from the case of individual decision making.

Second, in order to calculate the overall values, it is necessary to use multiattribute models. Since it is difficult to grasp the seriousness of impact categories like acidification, ozone layer depletion, ecotoxicity, and resource extraction in traditional LCA, and because the number of the categories (10 or more) is too large to be weighted (Goedkoop *et al*., 2000), life cycle impact assessment methodology has recently introduced both a damage assessment procedure and a weighting triangle (Hofstetter, 1998; Goedkoop and Spriensma, 2000). The latter is a method for systematically performing sensitivity analyses using a triangular graph. By contrast, the Eco-indicator 95 uses the so-called Distance-to-Target approach, in which the seriousness of an effect is related to the difference between the current and target values (Goedkoop, 1995).

This indicates the necessity of formally considering the multiattribute models in understanding the weights; that is, maximizing a multi-attribute value function will be one of the most general decision criteria and weight elicitation should be based on a sound measurement theory.

4.4 Concluding Remarks

The above discussion indicates that in order to evaluate health and environmental issues, it is necessary to rely on subjective value judgments (including weighting by a panel of experts). The line of argument developed already, however, cannot completely solve the following dilemma. On the one hand, cost-benefit analysis can be recognized as a method that is theoretically sound but yet is difficult to apply in many cases because placing monetary values on the outcomes of health and ecological issues is difficult and even immoral. On the other hand, multicriteria analysis might be expedient, although it sometimes necessitates important problem structuring steps.

Further consideration should be given to these kinds of issues, including the differences in the attitude toward solving the problems between multicriteria analysis (decision analysis) and economics.

References

Addiscott, T.M., 1996, Fertilizers and nitrate leaching, *Issues in Environmental Sciences* **5**:1–26.

Arondel, C., and Girardin, P., 1998, Sorting Cropping Systems on the Basis of Their Impact on Groundwater Quality, Technical Report 158, LAMSADE, Université Paris-Dauphine, Paris, France.

Audsley, E., ed., 1997, *Harmonisation of Environmental Life Cycle Assessment for Agriculture*, Silsoe Research Institute, Silsoe, UK.

Bailey, A.P., Rehman, T., Park, J., Keatinge, J.D.H., and Tranter, R.B., 1999, Towards a method for the economic evaluation of environmental indicators for UK integrated arable farming systems, *Agriculture, Ecosystems and Environment*, **72**(2):145–158.

Banxia Software, 1999, *Decision Explorer: User's Guide*, Banxia Software, Glasgow, UK.

Bertilsson, G., 1992, *Environmental Consequences of Different Farming Systems Using Good Agricultural Practices*, Proceedings No. 332, The Fertiliser Society, Peterborough, UK.

California Fertilizer Association, ed., 1995, *Western Fertilizer Handbook*, 8th ed., Interstate Publishers, Danville, IL, USA.

Croll, B.T., 1994, *Nitrate: Best Agricultural Practice for Water – The UK Experience*, Proceedings No. 359, The Fertiliser Society, Peterborough, UK.

Daily, G.C., Söderqvist, T., Aniyar, S., Arrow, K., Dasgupta, P., Ehrlich, P.R., Folke, C., Jansson, A., Jansson, B.-O., Kautsky, N., Levin, S., Lubchenco, J., Märler, K.-G., Simpson, D., Starrett, D., Tilman, D., and Walker, B., 2000, The value of nature and the nature of value, *Science*, **289**:395–396.

Dasbach, E., and Teutsch, S.M., 1996, Cost-utility analysis, in A.C. Haddix, S.M. Teutsch, P.A. Shaffer, and D.O. Dune, eds, *Prevention Effectiveness*, Oxford University Press, New York, Chapter 9, pp. 130–142.

Dyer, J.S., and Sarin, R.K., 1979, Measurable multiattribute value functions, *Operations Research*, **27**(4):810–822.

El-Swaify, S.A., and Yakowitz, D.S., eds, 1998, Multiple Objective Decision Making for Land, Water, and Environmental Management, in *Proceedings of the First International Conference on Multiple Objective Decision Support Systems (MODSS) for Land, Water, and Environmental Management: Concepts, Approaches, and Applications*, Lewis Publishers, Boca Raton, FL, USA.

Fischer, G.W., 1995, Range sensitivity of attribute weights in multiattribute value models, *Organizational Behavior and Human Decision Processes*, **62**(3):252–266.

Foltz, J.C., Lee, J.G., Martin, M.A., and Preckel, P.V., 1995, Multiattribute assessment of alternative cropping systems, *American Journal of Agricultural Economics*, **77**(2):408–420.

Fried, J.J., 1991, Nitrates and their control in the EEC aquatic environment, in I. Bogárdi and R.D. Kuzelka, eds, *Nitrate Contamination*, Springer-Verlag, Berlin, Germany, pp. 3–11.

Goedkoop, M., 1995, *The Eco-Indicator 95: Final Report*, PRé Consultants, Amersfoort, Netherlands.

Goedkoop, M., and Spriensma, R., 2000, *The Eco-Indicator 99: A Damage Oriented Method for Life Cycle Impact Assessment, Methodology Report*, 2nd ed., PRé Consultants, Amersfoort, Netherlands.

Goedkoop, M., Effting, S., and Collignon, M., 2000, *The Eco-Indicator 99: A Damage Oriented Method for Life Cycle Impact Assessment, Manual for Designers*, 2nd ed., PRé Consultants, Amersfoort, Netherlands.

Gold, M.R., Siegel, J.E., Russell, L.B., and Weinstein, C.M., 1996, *Cost-Effectiveness in Health and Medicine*, Oxford University Press, Oxford, UK.

Hayashi, K., 1999a, Multicriteria analysis for natural resource management: Coping with diversity of methods, in A. Musy, L.S. Pereira, and M. Fritsch, eds, *Emerging Technologies for Sustainable Land Use and Water Management*, Presses Polytechniques et Universitaires Romandes, Lausanne, Switzerland, pp. 10 (CD–ROM).

Hayashi, K., 1999b, Weighting in MODSS: Interpretation and effectiveness, Paper presented at the Second International Conference on Multiple Objective Decision Support Systems for Land, Water and Environmental Management, August 1–6, 1999, Brisbane, Australia.

Hayashi, K., 2000a, Evaluation of alternative farming systems by multicriteria analysis, *JARQ (Japan Agricultural Research Quarterly)*, **34**(3):209–213.

Hayashi, K., 2000b, Multicriteria analysis for agricultural resource management: A critical survey and future perspectives, *European Journal of Operational Research*, **122**(2):486–500.

Heathwaite, A.L., Burt, T.P., and Trudgill, S.T., 1993, Overview – the nitrate issue, in T.P. Burt, A.L. Heathwaite, and S.T. Trudgill, eds, *Nitrate: Processes, Patterns and Management*, John Wiley and Sons, New York, NY, USA, pp. 3–21.

Heijungs, R., ed., 1992a, *Environmental Life Cycle Assessment of Products: Backgrounds*, Center of Environmental Science, Leiden, Netherlands.

Heijungs, R., ed., 1992b, *Environmental Life Cycle Assessment of Products: Guide*, Center of Environmental Science, Leiden, Netherlands.

Heilman, P., Yakowitz, D.S., and Lane, L.J., 1997, Targeting farms to improve water quality, *Applied Mathematics and Computation*, **83**(2-3):173–194.

Hofstetter, P., 1998, *Perspectives in Life Cycle Impact Assessment*, Kluwer Academic Publishers, Dordrecht, Netherlands.

ISO 14042, 2000, *Environmental Management – Life Cycle Assessment – Life Cycle Impact Assessment*, International Standard Organization, Geneva, Switzerland.

Keeney, R.L., 1992, *Value-Focused Thinking: A Path to Creative Decisionmaking*, Harvard University Press, Cambridge, MA, USA.

Kelly, T.C., Lu, Y., and Teasdale, J., 1996, Economic-environmental tradeoffs among alternative crop rotations, *Agriculture, Ecosystems and Environment*, **60**(1):17–28.

Kumazawa, K., 1999, Present state of nitrate pollution in groundwater, *Japanese Journal of Soil Science and Plant Nutrition*, **70**(2):207–213. (in Japanese).

Lægreid, M., Bøckman, O.C., and Kaarstad, O., 1999, *Agriculture, Fertilizers and the Environment*, CABI Publishing, Wallingford, UK.

Lawrence, P.A., Stone, J.J., Heilman, P., and Lane, L.J., 1997, Using measured data and expert opinion in a multiple objective decision support system for semiarid rangelands, *Transactions in the ASAE*, **40**(6):1589–1597.

Malczewski, J., 1999, *GIS and Multicriteria Decision Analysis*, John Wiley and Sons, New York, NY, USA.

Matson, P.A., Parton, W.J., Power, A.G., and Swift, M.J., 1997, Agricultural intensification and ecosystem properties, *Science*, **277**:504–509.

MODSS'99, 1999, *Second International Conference on Multiple Objective Decision Support Systems for Land, Water and Environmental Management, Conference Program and Abstracts*, The State of Queensland, Department of Natural Resources, Brisbane, Australia.

Murray, C.J.L., 1994, Quantifying the burden of disease: the technical basis for disability-adjusted life years, *Bulletin of the World Health Organization*, **72**(3):429–445.

Nakanishi, J., and Gamo, M., 1998, Integrated method for evaluating the cancer and noncancer risks due to pesticides, *Reviews in Toxicology*, **2**(1–4):11–16.

OECD, 1997, *Environmental indicators for agriculture*, OECD, Paris, France.

Oka, T., Matsuda, H., and Kadono, Y., 1999, Ecological risk-benefit analysis of a wetland development based on risk assessment using ‘expected loss of biodiversity,’ Paper presented at Society for Risk Analysis, 1999 Annual Meeting, December 6, 1999, Atlanta, GA, USA.

Sleeswijk, A.W., Kleijn, R., van Zeijts, H., Reus, J.A.W.A., van Onna, M.J.G.M., Leneman, H., and Sengers, H.H.W.J.M., 1996, *Application of LCA to Agricultural Products*, Centre of Environmental Science, Leiden University, Leiden, Netherlands.

The Presidential/Congressional Commission on Risk Assessment and Risk Management, 1997a, *Framework for Environmental Health Risk Management*, Washington, DC, USA.

The Presidential/Congressional Commission on Risk Assessment and Risk Management, 1997b, *Risk Assessment and Risk Management in Regulatory Decision-Making*, Washington, DC, USA.

Tilman, D., 1999, Global environmental impacts of agricultural expansion: The need for sustainable and efficient practices, *Proceedings of the National Academy of Sciences*, **96**:5995–6000.

Tiwari, D.N., Loof, R., and Paudyal, G.N., 1999, Environmental-economic decision-making in lowland irrigated agriculture using multi-criteria analysis techniques, *Agricultural Systems*, **60**(1):99–112.

Vitousek, P.M., Aber, J.D., Howarth, R.W., Likens, G.E., Matson, P.A., Schindler, D.W., Schlesinger, W.H., and Tilman, D.G., 1997, Human alteration of the global nitrogen cycle: Sources and consequences, *Ecological Applications*, **7**(3):737–750.

von Nitzsch, R., and Weber, M., 1993, The effect of attribute ranges on weights in multiattribute utility measurements, *Management Science*, **39**(8):937–943.

Yakowitz, D.S., Lane, L.J., and Szidarovszky, F., 1993, Multi-attribute decision making: Dominance with respect to an importance order of the attributes, *Applied Mathematics and Computation*, **54**:167–181.

Chapter 5

Generating Efficient Alternatives in a Transboundary Air Pollution Negotiation Using Constraint Proposal Method

Pirja Heiskanen

Abstract

In this chapter we simulate a negotiation between Finland, Russia, and Estonia over the amounts of sulfur and nitrogen emissions in these countries. The chapter studies how the previously developed constraint proposal method can be used to help the negotiators find efficient agreements dominating the status quo solution of the negotiation. A simplification of the RAINS (Regional Acidification INformation and Simulation) model developed at IIASA is used to simulate the transportation of the air pollutants and to estimate their harmful effects to the environment. The decision makers' preferences are simulated using additive value functions where the essential criteria are the acidification and eutrophication effects of the depositions as well as the abatement costs. The numerical study indicates the success of the constraint proposal method in finding an individually rational efficient agreement with respect to the status quo solution.

Keywords: Negotiation, transboundary air pollution, multiple criteria, efficiency.

5.1 Introduction

During the last decades, the air pollution problem has been regarded as a major threat to the environment in Europe. The increased concentrations of sulfur and nitrogen oxides may affect plants, animals, human health, and materials. For example, many lakes in northern Europe have been slowly acidified due to the accumulated loading. The acidification is mainly caused by sulfur, nitrogen, and ammonia. In addition to causing acidification, excessive nitrogen deposition may cause unwanted eutrophication effects. The long-range transportation of pollutants across country borders makes air pollution a truly international problem and many international conferences on acidification have been held. Initially, the focus was on reducing acid rain by controlling sulfur emissions, but as these emissions dwindled, nitrogen-related effects have recently gained relative importance.

In this chapter we will study a transboundary air pollution problem between Finland, Russia, and Estonia. The countries negotiate over sulfur and nitrogen emissions in 2010. The aim of each country is to minimize its own abatement costs as well as the harmful effects of both acidification and eutrophication on the environment. We model the transportation of the air pollutants based on the RAINS (Regional Acidification INformation and Simulation) model developed at the International Institute for Applied Systems Analysis (IIASA). The harmful effects of the pollutants on the environment are modeled as exceedances over critical loads of different ecosystems. The abatement costs for both sulphur and nitrogen are estimated based on the data gathered in the Transboundary Air Pollution (TAP) project at IIASA and the value of environmental damage is modeled by a quadratic function.

In the numerical study, the previously developed constraint proposal method (Ehtamo *et al*., 1999; Heiskanen, 2001; Heiskanen *et al*., 2001) is used for searching efficient agreements dominating the status quo solution; that is, the emission levels induced by the current legislation in the countries in 2010. In the constraint proposal method, a neutral coordinator, called "mediator," gathers information on the parties' preferences by asking the parties to choose their optimal points on different sets of alternatives. Due to the interactive information gathering procedure the parties' value functions do not need to be explicitly revealed. This aspect makes the method particularly applicable in real life situations, where an explicit form for the value functions is not known or where the parties are not willing to reveal private information due to strategic reasons. In this preliminary study no real decision makers (DMs) are involved but the parties' answers are simulated using artificial value functions.

The search for efficient abatement scenarios in a transboundary air pollution problem between Finland, Russia, and Estonia has been considered previously by Kaitala *et al*. (1995) and by Kaitala and Pohjola (1999). In the former work, only sulfur abatements are considered and a joint optimal agreement is searched by minimizing the sum of the countries' value functions. The authors also discuss a possible need for financial transfers in making the solution individually rational. In Kaitala and Pohjola (1999), local information on the countries' cost and damage functions is used to establish a gradual emission abatement program that works toward the joint optimal levels.

5.2 Model for Transboundary Air Pollution

In our model, Finland, Estonia, and Russia negotiate over the amounts of sulfur and nitrogen emissions in 2010. We consider only those parts of Russia that are close to Finland and Estonia and thus affect depositions in those countries. Thus, four regions are considered in the model, namely, Finland, the Kola and Karelia region, the St. Petersburg region, and Estonia. In the sequel, vectors $x^S \in \mathbb{R}^4$ and $x^N \in \mathbb{R}^4$ denote the sulfur (SO_2) and the nitrogen (NO_x) emissions of the regions. The decision vector is $x := ((x^S)^T \ (x^N)^T)^T \in \mathbb{R}^8$, where

x_1 := the sulfur emission of Finland,
x_2 := the sulfur emission of the Kola and Karelia region,
x_3 := the sulfur emission of the St. Petersburg region,
x_4 := the sulfur emission of Estonia,
x_5 := the nitrogen emission of Finland,
x_6 := the nitrogen emission of the Kola and Karelia region,
x_7 := the nitrogen emission of the St. Petersburg region,
x_8 := the nitrogen emission of Estonia,

and all the emissions are given in kt/year.

We assume that if no negotiated agreement is reached, each country carries out the abatements required by current legislation. Thus a forecast for the emission levels in 2010 obtained using the current legislation scenario (clbunn scenario in Rains 7.2 software) can be considered as a status quo solution of the negotiation. The numerical value for it is $\bar{x} = (155\ 473\ 136\ 175\ 152\ 86\ 170\ 73)^T$.

5.2.1 The transportation model

The transportation model for air pollutants simulates the atmospheric processes and deposition rates of different pollutants. In RAINS, the transportation model is a linear mapping from the regions to a 150×150 km^2 grid covering all of Europe. The transfer matrices used in the model are provided by the EMEP (European Monitoring and Evaluation Programme) Meteorological Synthesizing Center-West (MSC-W) at the Norwegian Meteorological Institute. Let q_j^S, q_j^N denote the sulfur and nitrogen depositions in the grid cell j, $j = 1, \dots, n$ (n is the number of the grid cells). Then the total depositions in the countries, $\bar{q}^S \in I\!R^3$ and $\bar{q}^N \in I\!R^3$, are obtained as a weighted sum of the depositions in the grid cells. Here the weight for a grid cell is the country's area in that grid cell (for details see Heiskanen, 2000). The transportation model for the emissions of the regions to the depositions in the countries is the following:

$$\bar{q}^S = G^S x^S + b^S \tag{5.1a}$$

$$\bar{q}^N = G^N x^N + b^N \; . \tag{5.1b}$$

Here $G^S \in I\!R^{3\times 4}$ and $G^N \in I\!R^{3\times 4}$ are the transportation matrices for sulfur and nitrogen, respectively. Vectors $b^S \in I\!R^3$ and $b^N \in I\!R^3$ contain depositions due to natural background depositions and emissions originating outside the region of interest. The background depositions and the emissions for other countries than Finland, Russia, and Estonia are obtained using the current legislation scenario in the RAINS model. All the depositions are given in kt/year. The countries are enumerated so that the first components of the vectors denote the depositions in Finland, the second components denote the depositions in Estonia, and the last components denote the depositions in Russia. The numerical values for the transportation matrices and for the background depositions are as follows:

$$G^S = \begin{pmatrix} 0.133 & 0.040 & 0.026 & 0.038 \\ 0.013 & 0.001 & 0.006 & 0.045 \\ 0.098 & 0.178 & 0.285 & 0.194 \end{pmatrix}$$

$$G^N = \begin{pmatrix} 0.044 & 0.010 & 0.014 & 0.021 \\ 0.004 & 0.000 & 0.003 & 0.012 \\ 0.079 & 0.064 & 0.124 & 0.073 \end{pmatrix}$$

$$b^S = (40.99 \; 16.29 \; 943.22)^T$$
$$b^N = (34.01 \; 10.47 \; 593.64)^T$$

5.2.2 Critical loads and exceedance functions

The effect of a pollutant on the environment depends not only on the amount of deposition, but also on the sensitivity of the ecosystem. For example, a sulfur level that is harmless to an ecosystem in central Europe may already pose a serious threat to a sensitive ecosystem in Lapland. Therefore, the concept of *critical load* (CL) has been introduced. It is defined as "a quantitative estimate of an exposure to one or more pollutants below which significant harmful effects on specified sensitive elements of the environment do not occur according to present knowledge" (Nilsson and Grennfelt, 1988). The critical loads are usually defined separately for every ecosystem, and, on their basis an average critical load for every grid cell is obtained.

The amount of the harmful effect on the environment is measured by exceedance function (see, e.g., Nilsson and Grennfelt, 1988). If only one pollutant contributes to an effect, e.g., nitrogen to eutrophication, a unique critical load can be calculated and compared with deposition. Therefore, if the critical load of nutrient nitrogen for a grid cell j is denoted by CL_j^{nut}, then the exceedance for eutrophication for that cell is obtained as

$$e_j^{nut} = q_j^N - CL_j^{nut} \; . \tag{5.2}$$

In the case of acidification, the definitions for critical load and exceedance are more complicated, since both sulfur and nitrogen contribute to acidification. However, one equivalent of sulfur contributes more to excess acidity than one equivalent of nitrogen. As long as the deposition of nitrogen stays below the minimum critical load of nitrogen CL^{Nmin}, all deposited nitrogen is consumed by sinks. Then the critical load for sulfur is given by CL^{Smax}. The maximal critical load for nitrogen is CL^{Nmax} in the case of no sulfur deposition. Otherwise the critical load for nitrogen is between the maximal and minimal value. Here we approximate the critical load function by a linear function. Then the exceedance for acidification for a grid cell j gets the following form (for details see Heiskanen, 2000):

$$e_j^{acid} = q_j^S + [\frac{CL_j^{Smax}}{CL_j^{Nmax} - CL_j^{Nmin}}][q_j^N - CL_j^{Nmax}] \; . \tag{5.3}$$

The *accumulated exceedance* for a country is a weighted sum over the exceedances of the grid cells. Here the weight of a grid cell is the fraction of area of the country in that cell. To obtain a measure that is independent of the area of the country, *average accumulated exceedance* is often used instead of the accumulated exceedance. It is obtained by dividing the accumulated exceedance by the total area of the country. The average accumulated exceedance has the same dimension (mg/m^2/year) as deposition, and thus they can be directly compared.

Combining the previous results, we obtain a linear mapping from the emissions of the regions to the average accumulated exceedances of the countries as follows

$$E^{acid}(x^S, x^N) = T^1 x^S + T^2 x^N + v^1 \tag{5.4a}$$

$$E^{eut}(x^N) = T^3 x^N + v^2 \ , \tag{5.4b}$$

where the numerical values for matrices T^1, T^2, T^3 and vectors v^1 and v^2 are the following:

$$T^1 = \begin{pmatrix} 0.398 & 0.119 & 0.079 & 0.114 \\ 0.286 & 0.014 & 0.131 & 0.979 \\ 0.028 & 0.051 & 0.082 & 0.056 \end{pmatrix}$$

$$T^2 = \begin{pmatrix} 0.088 & 0.021 & 0.028 & 0.043 \\ 0.097 & 0.001 & 0.069 & 0.295 \\ 0.022 & 0.018 & 0.035 & 0.021 \end{pmatrix}$$

$$T^3 = \begin{pmatrix} 0.132 & 0.030 & 0.04 & 0.062 \\ 0.093 & 0.001 & 0.065 & 0.273 \\ 0.023 & 0.019 & 0.036 & 0.021 \end{pmatrix}$$

$$v^1 = (\text{-228 -2566 -636})^T$$
$$v^2 = (\text{-93 -55 -135})^T$$

5.3 Objective Functions

When searching for an abatement scenario, each country is interested in minimizing its own abatement costs as well as the harmful effects of the pollutants. Therefore, there are three criteria to be considered: abatement costs, acidification damage, and eutrophication damage.

5.3.1 Abatement costs

The costs for reducing the emissions in different regions are estimated based on the data obtained from the RAINS model. The RAINS model uses the information on available technologies and their costs, as well as the types of fuels and the estimated energy consumptions in the regions, to construct piecewise linear cost curves for the regions (Cofala and Syri, 1998a,b). Here we use the data to estimate quadratic cost curves for sulfur and nitrogen. Kaitala and Pohjola (1999) use the same quadratic

Table 5.1. The parameters of the cost functions.

Region (*i*)	α_i^S	β_i^S	γ_i^S	α_i^N	β_i^N	γ_i^N
Finland	0.0110	-5.1140	701.2	0.0281	-12.8541	1796.0
Kola and Karelia	0.0003	-0.6058	234.9	0.0116	-3.0053	161.1
St. Petersburg	0.0009	-0.6494	86.8	0.0056	-2.8621	303.9
Estonia	0.0008	-0.5829	78.5	0.0154	-3.3946	158.1

form for the cost functions. The parameters of the functions are estimated from the data using the ordinary least squares method. The abatement costs for sulfur and nitrogen in region i are denoted by $c_i^S(x_i^S)$ and $c_i^N(x_i^N)$, respectively, and they are of the form

$$\begin{aligned} c_i^S(x_i^S) &= \alpha_i^S(x_i^S)^2 + \beta_i^S x_i^S + \gamma_i^S \\ c_i^N(x_i^N) &= \alpha_i^N(x_i^N)^2 + \beta_i^N x_i^N + \gamma_i^N. \end{aligned} \tag{5.5}$$

The costs are given in M euro/year and the emissions are given in kt/year. The values of the parameters for different regions are given in *Table 5.1*. *Figure 5.1* and *Figure 5.2* show the original cost curves for year 2010 obtained from the RAINS model using the default scenario (no emission control), as well as the estimated quadratic functions.

5.3.2 Damage functions

The value of the harmful effects of the pollutants on the environment is measured by damage functions. In the sequel, the value of acidification damage for country k is denoted by $D_k^{acid}(x^S, x^N)$ and the value of eutrophication damage is denoted by $D_k^{eut}(x^N)$. The value of acidification damage is assumed to be a quadratic function of the average accumulated exceedance. The function is scaled so that the damage is zero if the emissions are zero. So, the value of damage caused by the depositions in country k is $\frac{1}{2}[E_k^{acid}(x^S, x^N) - E_k^{acid}(0,0)]^2$, where E_k^{acid} is given by Equation (5.4a). We assume that a country is also concerned about acidification in neighboring countries, although not as much as about that within its own borders. The parameter μ defines how valuable a country considers the damage in surrounding countries compared to the damage in its own country. For example, the value $\mu = 0.1$ means that if acidification is at the same level everywhere, it is ten times more important to a country to decrease acidification within its own boundaries than elsewhere. On the other hand, if $\mu = 0.0$, a country is concerned only about the environmental damage within its own borders. The value of eutrophication damage is modeled in the same way as the value of acidification damage. Thus the damage functions are of the form:

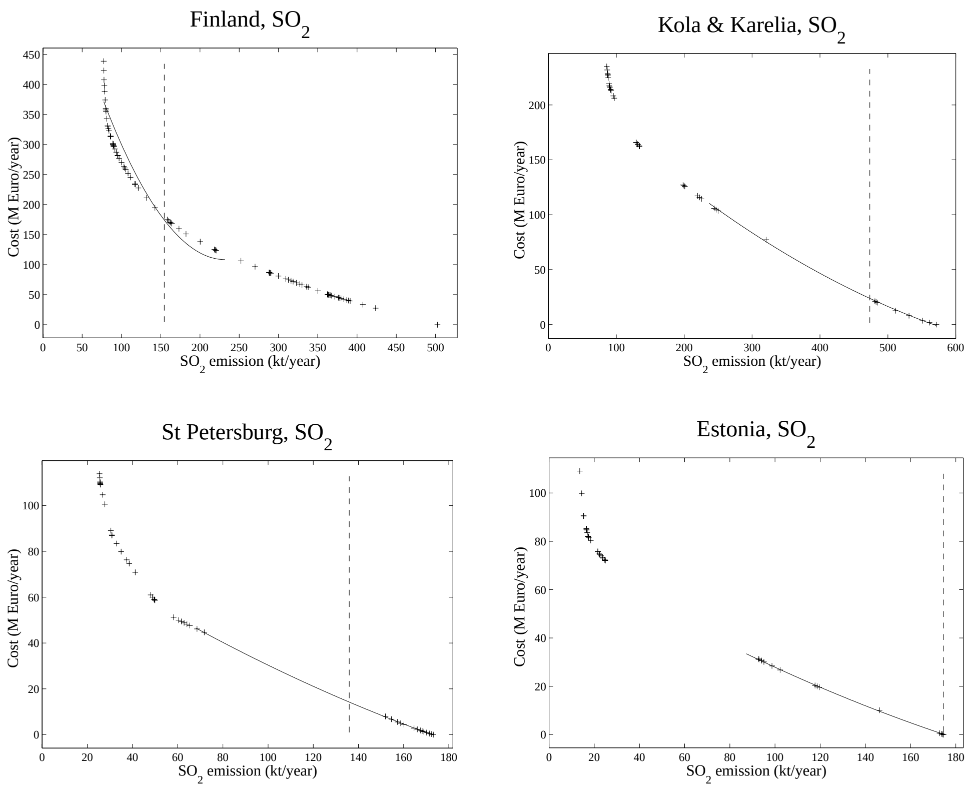


Figure 5.1. The original data (+) and the estimated yearly abatement costs (solid line) for SO_2 in different regions. The status-quo emission level is indicated by a dashed vertical line.

$$\begin{aligned} D_k^{acid}(x^S, x^N) &= \tfrac{1}{2}\{[E_k^{acid}(x^S, x^N) - E_k^{acid}(0,0)]^2 \\ &+ \mu \textstyle\sum_{i \neq k} [E_i^{acid}(x^S, x^N) - E_i^{acid}(0,0)]^2\} \end{aligned} \tag{5.6a}$$

$$\begin{aligned} D_k^{eut}(x^N) &= \tfrac{1}{2}\{[E_k^{eut}(x^N) - E_k^{eut}(0)]^2 + \mu \textstyle\sum_{i \neq k} [E_i^{eut}(x^N) \\ &- E_i^{eut}(0)]^2\} \end{aligned} \tag{5.6b}$$

5.3.3 Estimating the value functions

Here we estimate value functions for the three countries. In this preliminary study, artificial value functions are used to simulate the parties' answers in the constraint proposal method. In practice, the preference information would be obtained directly from the DMs, and the elicitation of the value functions would not be needed.

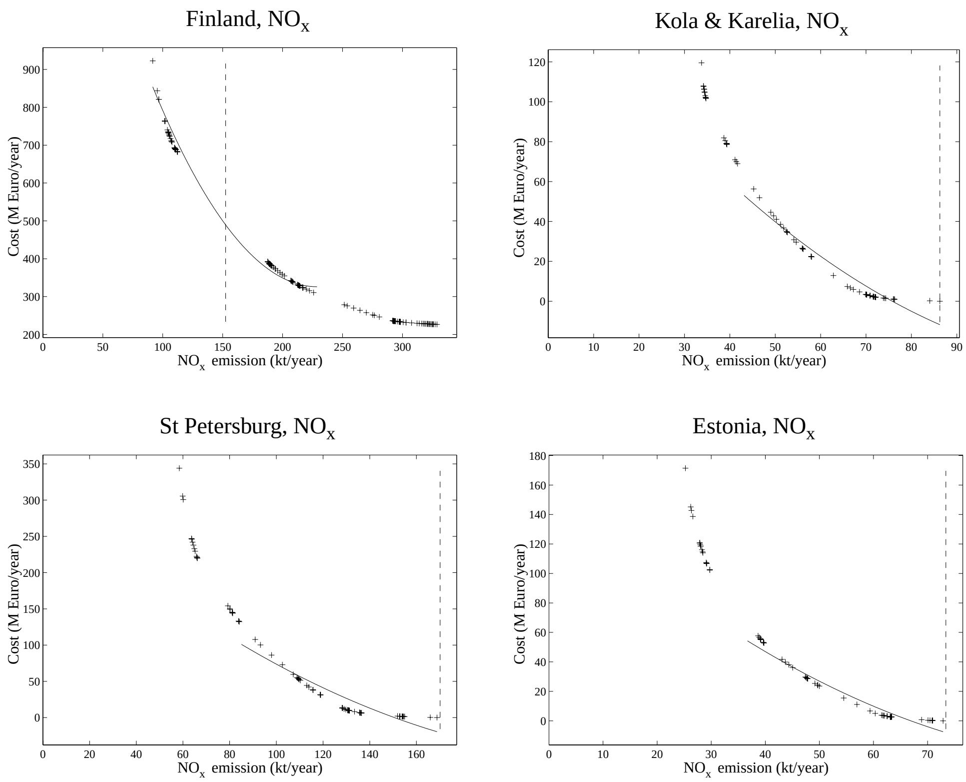


Figure 5.2. The original data (+) and the estimated yearly abatement costs (solid line) for NO_x in different regions. The status-quo emission level is indicated by a dashed vertical line.

We assume that the value functions are additive with respect to the three criteria, namely, the abatement costs, acidification damage, and eutrophication damage. Therefore the value functions take the following form:

$$u_{FIN}(x^S, x^N) = c_1^S(x_1^S) + c_1^N(x_1^N) + \pi_1 D_1^{acid}(x^S, x^N) + \eta_1 D_1^{eut}(x^N) \tag{5.7a}$$

$$u_{EST}(x^S, x^N) = c_4^S(x_4^S) + c_4^N(x_4^N) + \pi_2 D_2^{acid}(x^S, x^N) + \eta_2 D_2^{eut}(x^N) \tag{5.7b}$$

$$\begin{aligned} u_{RUS}(x^S, x^N \quad &= \quad c_2^S(x_2^S) + c_2^N(x_2^N) + c_3^S(x_3^S) + c_3^N(x_3^N) \\ &+ \quad \pi_3 D_3^{acid}(x^S, x^N) + \eta_3 D_3^{eut}(x^N) \end{aligned} \tag{5.7c}$$

Here π_i and η_i, $i = 1, \ldots, 3$, define the monetary value of acidification and eutrophication damages, respectively. For example, a value $\pi_1 = 0.01$ means that, in Finland, a decrease in costs by one million euro/year is considered as valuable as a one-hundred unit decrease in the acidification damage. The values for the parameters are estimated by using a so called revealed preference approach suggested by Mäler (1990). In this approach, the values are chosen so that the current legislation scenario $\bar{x}$ is a Nash solution of the problem (see also Kaitala and Pohjola, 1999). In the current situation it is not possible to choose the values of parameters π and η so that $\bar{x}$ would be a Nash solution, because Russia controls four decision variables while there are only two parameters to be estimated in her value function. Hence, we choose the values for the parameters so that the Nash solution is close to $\bar{x}$ but its components are greater than or equal to $\bar{x}$. While estimating π and η, the value of parameter μ was 0.00, which means that the countries are not concerned about the environmental problems in the neighboring countries. The numerical values obtained for π and η are

$$\begin{aligned} \pi &= (2.42 \cdot 10^{-2} \;\; 1.11 \cdot 10^{-3} \;\; 7.87 \cdot 10^{-2})^T \\ \eta &= (8.42 \cdot 10^{-1} \;\; 8.33 \cdot 10^{-2} \;\; 1.64)^T \end{aligned}$$

and they are given in M euro (m^2/mg)^2. Then the Nash solution of the problem is $x^{Nash} = (155\ 510\ 136\ 175\ 152\ 108\ 170\ 73)^T$.

5.4 Searching for Pareto-Optimal Solutions

In this section we demonstrate how the constraint proposal method can be used to find Pareto-optimal agreements in the previously described transboundary air pollution problem. The idea is to search for Pareto-optimal agreements dominating the status-quo solution of the negotiation; namely, the emission levels obtained by applying the abatements required by current legislation. Consequently, we restrict the decision set so that the emissions are not allowed to be greater than the emissions at the status-quo solution $\bar{x}$. Since the cost functions were estimated using only values where emissions are at least half of $\bar{x}$, we use $\frac{1}{2}\bar{x}$ as a lower limit. Thus the decision set is $X = \{x \in \mathbb{R}^8 | \ \frac{1}{2}\bar{x} \leq x \leq \bar{x}\}$, where the numerical value for $\bar{x}$ is $\bar{x} = (155\ 473\ 136\ 175\ 152\ 86\ 170\ 73)^T$.

If not stated otherwise, the value of μ is 0.05 in the numerical examples. This means that we assume the countries to be somewhat concerned about environmental damage in the neighboring countries. Under this assumption, the countries' optimal emission vectors on decision set X are as follows:

$$\bar{x}^{FIN} = (155 \quad 237 \quad 68 \quad 87 \quad 152 \quad 43 \quad 85 \quad 37)^T \tag{5.8a}$$

$$\bar{x}^{EST} = (77 \quad 237 \quad 68 \quad 175 \quad 76 \quad 43 \quad 85 \quad 73)^T \tag{5.8b}$$

$$\bar{x}^{RUS} = (77 \quad 473 \quad 118 \quad 87 \quad 76 \quad 86 \quad 165 \quad 37)^T \tag{5.8c}$$

In the optimal solution for each country, the emissions of neighboring countries are at their lowest level. However, since μ is now strictly positive, each country's own emissions are not always at their highest level.

5.4.1 Constraint proposal method

In the constraint proposal method, candidates for Pareto-optimal solutions are searched for by iterating with a set of linear constraints going through a fixed reference point r in the decision set. At every iteration, the countries indicate their optimal solutions on the intersection of the constraints and the decision set. Each country optimizes with respect to the whole decision vector $x \in \mathbb{R}^8$, and not only with respect to her own emissions. If the countries' optimal solutions coincide, a candidate for a Pareto-optimal solution has been found. Otherwise the mediator, who works as a neutral coordinator, updates the set of linear constraints and the iteration is continued.

In the constraint proposal method, the choice of the reference point r is very important, since it determines the solution. The solution corresponding to a reference point is not necessarily unique; there may be several distinct solutions. However, the solutions are always individually rational with respect to the reference point, which is a desirable property of a negotiation support system (Heiskanen, 2001). In the constraint proposal method, several different Pareto-optimal solutions can be systematically generated by varying the reference point appropriately. For a comprehensive description of the method and its properties, please see Heiskanen (2001). Before applying the constraint proposal method here, the problem is scaled so that the decision variables obtain values between 0 and 1. The scaling is used to stabilize the computation and it does not affect the formulation of the DMs' problems. A detailed description of the scaling is given in Heiskanen (2000).

5.4.2 An efficient solution dominating the status quo

First we generate one Pareto-optimal solution dominating the status-quo solution. By generating a feasible solution dominating the status-quo solution, one can show the existence of joint gains and the usefulness of the negotiation. Since the solution of the constraint proposal method is individually rational with respect to the reference point, a Pareto-optimal solution dominating the current legislation scenario can be obtained easily by using the status quo solution as a reference point.

Table 5.2. Values of the countries' value functions.

k	$u_k(\bar{x})$	$u_k(s)$	$u_k(\bar{x}^k)$	Gain (%)
FIN	1,620	1,415	1,200	49
EST	128	102	70	45
RUS	644	494	334	48

Table 5.3. Total depositions of SO_2 and NO_x in the countries at the solution point, at the status-quo solution, and at the countries' individual optima.

	Deposition at $\bar{x}$		Deposition at s		Deposition at $\bar{x}^k$	
k	SO_2	NO_x	SO_2	NO_x	SO_2	NO_x
FIN	91	46	75	42	76	46
EST	27	13	23	12	26	13
RUS	1,115	638	1,043	623	1,086	651

In the numerical heuristics of the constraint proposal method a value $\epsilon = 0.1$ was used as a stopping criterion. The number of iterations needed to achieve the required accuracy was 7. The solution that averages the countries' optimal solutions on the final affine set is $s = (131\ 273\ 70\ 93\ 108\ 86\ 87\ 61)^T$. At the solution, the sulfur emissions of the Kola and Karelia region and Estonia are almost at their lowest level, as is the nitrogen emission of the St. Petersburg region. *Table 5.2* shows the countries' values at the status quo solution $\bar{x}$, at the solution s, and at their individual optima on the decision set X. We also calculated the percentage of gain achieved by moving from the status-quo solution to the solution point (compared to moving to the individual optimum); that is, the gain is calculated as

$$gain := \frac{u_k(s) - u_k(\bar{x})}{u_k(\bar{x}^k) - u_k(\bar{x})} . \tag{5.9}$$

One should note that all the countries cannot obtain 100% gain simultaneously since their optima on X do not coincide.

Table 5.2 shows that each country achieves a considerable gain by moving to the solution point from the status-quo solution. The relatively small change seen in the countries' value functions results from the fact that a considerable amount of the pollution deposited in the three countries originates from outside the regions considered here and thus cannot be affected by the actions of these countries. Reducing the emissions from the status-quo level to the level corresponding the solution is of course costly to the countries. In this case, the additional abatement costs for Finland, Estonia, and Russia are 292, 47, and 221 M euro/year, respectively. *Table 5.3* shows the corresponding sulfur and nitrogen depositions in the countries.

Table 5.4. The effect of parameter μ on the emission levels of the solution.

		SO_2				NO_x			
μ	Iterations	FIN	K&K	STP	EST	FIN	K&K	STP	EST
0.01	16	148	388	98	96	115	86	115	63
0.05	7	131	273	70	93	108	86	87	61
0.10	8	133	237	68	94	109	53	85	70

5.4.3 The effect of μ

Since the value of μ indicates how concerned the DMs are about the environmental effects in neighboring countries, we want to see how it affects the solution obtained. We varied the value of μ and for each value of it, we generated a Pareto-optimal solution dominating the status-quo solution. The values of parameters π and η were not re-estimated. The results are shown in *Table 5.4*. One can see that when the countries become more concerned about environmental damage in their neighbor's territories, the emissions in Russia should especially be further reduced.

5.4.4 Approximation to the Pareto frontier

The negotiating parties are not always satisfied with one candidate for an agreement, but want to explore all possible solutions and then negotiate over them. This is understandable, since different solutions divide the additional gain in different ratios between the parties. Therefore we will generate here an approximation to the whole Pareto-optimal frontier. The approximation is obtained by generating Pareto-optimal solutions starting from six different reference points that are convex combinations of the DMs' optima on the decision set. In a two-DM case, this way of choosing the reference point guarantees the Pareto-optimality of solutions. Numerical examples show that most of the solutions are Pareto-optimal even in the case where several parties are negotiating (see Section 3.3 in Heiskanen, 2001).

The reference points, solutions, and number of iterations required are shown in *Table 5.5*. The number of iterations needed varied from 8 to 14, the average being 10.83. At most of the generated solutions, sulfur emissions were considerably reduced at the Kola and Karelia region, the St. Petersburg region, and Estonia. On the other hand, the nitrogen emission of Kola and Karelia was always almost at its highest level. The approximation for the Pareto-optimal frontier made here is, of course, very rough. However, we did not want to make a more accurate approximation since in practice it is very difficult and laborious to obtain preference information from the DMs.

In *Table 5.6* we show the values of the countries' value functions at the reference points and at the solution points given in *Table 5.5*. We also calculated the

Table 5.5. Reference points (r^i), corresponding solutions (s^i) and iterations needed when an approximation to the Pareto frontier was generated.

	SO_2				NO_x				
	FIN	K&K	STP	EST	FIN	K&K	STP	EST	Iterations
r^1	124	284	78	105	122	52	101	44	14
s^1	137	266	72	87	121	86	86	49	
r^2	108	331	88	105	107	60	117	44	8
s^2	118	354	71	87	111	86	104	51	
r^3	93	379	98	105	91	69	133	44	12
s^3	103	364	76	87	91	85	122	47	
r^4	108	284	78	122	107	52	101	51	10
s^4	132	282	68	96	107	84	87	65	
r^5	93	331	88	122	91	60	117	51	10
s^5	108	360	74	87	92	86	94	62	
r^6	93	284	78	140	91	52	101	59	11
s^6	111	285	68	121	90	84	86	72	

Table 5.6. The values of the countries' value functions at the reference points (r) and at the solution points (s).

k	$u_k(r^1)$	$u_k(s^1)$	Gain (%)	Cost (M euro/yr)	$u_k(r^2)$	$u_k(s^2)$	Gain (%)	Cost (M euro/yr)
FIN	1,349	1,340	67	196	1,463	1,444	42	304
EST	123	118	17	69	121	117	19	67
RUS	510	490	50	223	456	442	65	157
k	$u_k(r^3)$	$u_k(s^3)$	Gain (%)	Cost (M euro/yr)	$u_k(r^4)$	$u_k(s^4)$	Gain (%)	Cost (M euro/yr)
FIN	1,596	1,568	12	492	1,445	1,426	46	294
EST	119	118	17	74	105	99	51	39
RUS	409	397	80	123	515	503	45	220
k	$u_k(r^5)$	$u_k(s^5)$	Gain (%)	Cost (M euro/yr)	$u_k(r^6)$	$u_k(s^6)$	Gain (%)	Cost (M euro/yr)
FIN	1,578	1,549	17	465	1,561	1,554	16	473
EST	104	99	51	48	91	84	76	21
RUS	460	443	65	171	521	509	44	221

percentage of gain achieved by moving from the status-quo solution $\bar{x}$ to the solution $s := s^i$ calculated by Equation (5.9) as well as the additional abatement costs (M euro/year) for reducing emissions from the status-quo level.

Table 5.6 shows that each country gains by moving from the status-quo solution to any of the solutions. Depending on the solution, the gain for Finland varies between 12% and 67%, the gain for Estonia varies between 17% and 76%, and the gain for Russia varies between 44% and 80%. One can also see that Finland almost always pays most of the emission abatement costs, even though emissions are reduced much more elsewhere. This situation results from the high marginal costs for emission abatements in Finland, observable in *Figure 5.1* and *Figure 5.2*.

5.5 Discussion

In this chapter we tested the constraint proposal method with realistic data. Both the model and the problem settings were real and the value functions were estimated on the basis of the data. We assumed the DMs' answers to be rational and accurate, which, of course, may not be the case in the real world.

The numerical study indicates the success of the constraint proposal method in finding efficient agreements dominating the status-quo solution of the negotiation. All the solutions were individually rational and considerable joint gains were achieved by moving to the solution from the status quo. At most of the generated solutions, the sulfur emissions of the Kola and Karelia region, the St. Petersburg region, and Estonia were considerably reduced. On the other hand, the nitrogen emission of Kola and Karelia was at its highest level. This indicates that, starting from the status-quo solution, the reductions in sulfur emissions are more important than those of nitrogen. If the countries become more concerned about the environmental damage in neighboring countries, emissions in Russia should be further reduced.

These results are important, since one must always evaluate models for decision support carefully before approaching the true DMs. Preliminary tests help point out possible shortcomings and improve the method before the real DMs, whose time and expertise are valuable resources, are involved in the process.

Acknowledgments

I want to thank Professors Markku Kallio and Pekka Korhonen for their idea to study the transboundary air pollution problem as well as Professor Seppo Salo for his comments. I am also grateful to the personnel of the Transboundary Air Pollution project at IIASA for clarifying discussions and for providing me with the RAINS software. This work has been financially supported by grants from the Ella and Georg Ehrnrooth's Foundation, the KAUTE Foundation, Yrjö Jahnsson Foundation, and the Foundation of the Helsinki School of Economics.

References

Cofala, J., and Syri, S., 1998a, Nitrogen oxides emissions, abatement technologies and related costs for Europe in the RAINS model database, IR-98-88, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Cofala, J., and Syri, S., 1998b, Sulfur emissions, abatement technologies and related costs for Europe in the RAINS model database, IR-98-35, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Ehtamo, H., Hämäläinen, R.P., Heiskanen, P., Teich, J.E., Verkama, M., and Zionts, S., 1999, Generating Pareto solutions in a two-party setting: Constraint proposal methods, *Management Science*, **45**(12):1697–1709.

Heiskanen, P., 2001, Generating Pareto-optimal boundary points in multi-party negotiations using constraint proposal method, *Naval Research Logistics*, **48**:1–16.

Heiskanen, P., 2000, *Application of constraint proposal method in transboundary air pollution negotiations*, Working Paper W-244, Helsinki School of Economics, Helsinki, Finland.

Heiskanen, P., Ehtamo, H., and Hämäläinen, R.P., 2001, Constraint proposal method for computing Pareto solutions in *n*-party negotiations, *European Journal of Operational Research*, **133**(1):44–61.

Kaitala, V., Mäler, K.-G., and Tulkens, H., 1995, The acid rain game as resource allocation process with an application to the international cooperation among Finland, Russia and Estonia, *Scandinavian Journal of Economics*, **97**(2):325–343.

Kaitala, V., and Pohjola, M., 1999, Acid rain and environmental aid: Transboundary air pollution between Finland, Russia and Estonia, in N. Hanley and H. Folmer, eds, *Game Theory and the Environment*, Eduard Elgar, Aldershot, UK.

Mäler, K.-G., 1990, International economic problems, *Oxford Review of Economic Policy*, **6**:80–108.

Nilsson, J., and Grennfelt, P., 1988, *Critical Loads for Sulphur and Nitrogen*, Nordic Council of Ministers, Vol. 97 of *Nord*, Copenhagen, Denmark.

Chapter 6

A Multi-Criteria Analysis for Open Space Conservation in New York State

Sigrid Stagl, Graham Cox, Jon Erickson, and Klaus Hubacek

Abstract

Pursuing the goal of large-scale ecosystem protection, the State of New York has for decades been acquiring private land parcels in the Adirondack State Park. While effective in terms of environmental protection, the process has repeatedly caused tensions with local communities who found themselves deprived of development possibilities. To ease these tensions, the involved parties agreed that a more open and participatory process was needed for guiding the Park's future development and conservation strategies. To push further the improvements implemented in 1998 with the State Open Space Conservation Plan, this paper suggests a framework for ranking alternative projects by use of multiple criteria decision aid (MCDA). With reference to data from parcels acquired by the State in the past, it is shown how MCDA is able to take into account a number of (in part conflicting) goals in a coherent and transparent way. For the case study, the NAIADE method was chosen, which can handle a number of different types of data and which supports the analysis of the structure of power interests and stakeholders by means of an institutional analysis.

Keywords: multiple criteria decision aid, NAIADE, sustainable land use, environmental policy, United States.

6.1 Introduction

The Adirondack Park, located in Northern New York State of the United States, is a unique combination of public lands protected by the State Constitution as "forever wild forest" and privately owned land regulated by state and local zoning laws. The State land includes roughly 47 percent of the Park. This combination of ownership has created an unprecedented application of land-use planning compatible with large-scale ecosystem protection. Although this complex pattern of public and private ownership has developed over the past century more by chance than design, New York has to a large degree been able to protect the ecological integrity of the largest park in the contiguous United States (Erickson, 1998).

These gains in environmental protection on behalf of the State's population at large were achieved at the expense of individual development rights of private landowners and local communities. Conflict between public agencies, local and statewide non-governmental organizations, and citizens has erupted around most new State land acquisitions, policy proposals, or management. One reason for these tensions is that today's Park has evolved from a rather top-down acquisition and planning process with little input from local communities. The 105 towns and villages within the Park boundary were left to bear the burden of real or perceived conflicts between a state agenda of environmental protection and a local agenda of economic development (Erickson and O'Hara, 2000).

Given the tensions, controversies, and political shakeup that resulted from the Adirondack Park's top-down protection efforts, decision-makers have agreed that a more open and participatory process is needed for guiding the Park's future development and conservation strategies. This has been a motivating force during the last decade of statewide initiative to devise a more acceptable and transparent land acquisition and management strategy and process. As part of the new statewide process, the relevant public agencies developed a system to evaluate and justify parcel acquisition using diverse criteria and given limited annual budgets. This system is also part of a process that has attempted to elicit public participation and communicate the rationale for continued statewide open space acquisition.

Very few acquisitions in the Adirondacks have occurred under this new formal system, however, numerous projects are currently entering the evaluation process. This paper will review the process of State land acquisition and report on ways to improve the ranking and decision-making processes by applying multiple criteria decision aid (MCDA).

MCDA can be used to support decision making in cases where conflicting economic, environmental, societal, institutional, technical, and aesthetic objectives may be involved. This multidimensionality is characteristic of most questions concerning sustainable development. MCDA allows for the use of heterogeneous criteria such as costs and benefits of the project, environmental quality in physical and qualitative terms, social impact in non-monetary terms, and even verbal descriptions of aesthetics.

6.2 The Case of the Adirondack Park

Decisions on State land acquisition in the Adirondacks have been characterized as a top-down process. At their most extreme, land purchases in the 1970s and 1980s were at times seen as projects stemming from the personal agendas or "wish lists" of state conservation officers or politicians. Opportunity for public comment or local consultation was rarely, if ever, encouraged. The public increasingly felt disconnected from both the rationale and process of spending taxpayer money, principally from State bonds approved by voters, on new acquisitions. State land in the Adirondacks had the additional burden on taxpayers of payment of property taxes in perpetuity.

In 1990, for the first time in State history, New York voters failed to pass an environmental bond issue that would have provided funds for significant additions to statewide holdings, most significantly in the Adirondack Park. Many have pointed to this event as the turning point in state open space planning and acquisition. Change, at least in spirit, that created a transparent process, clear rationale, and proposed management for State property was needed in order to instill faith in and restore finances for new acquisitions.

The *New York State Open Space Plan* of 1998 was the result of these events. It represents New York's first comprehensive plan and justification for statewide open space protection through land acquisition and conservation easements. Most significantly, the plan outlines a formal process for project evaluation and review. Any project under consideration for State land protection must now pass through six screens before the Commissioner of the Department of Environmental Conservation (DEC) will consider purchase of the property in fee or easement (i.e., purchase of development rights only). The six screens are outlined in *Figure 6.1*. Any person or private or public organization can propose a parcel of land to the State for protection. The State first and foremost tries to work only with a willing seller, reserving powers of eminent domain for rare circumstances. Starting with the "Resource Area Screen," the appropriate regional office of the DEC or the Office of Parks, Recreation, and Historic Preservation (OPRHP) determines whether a proposed parcel falls into either a resource area or linear system targeted in the

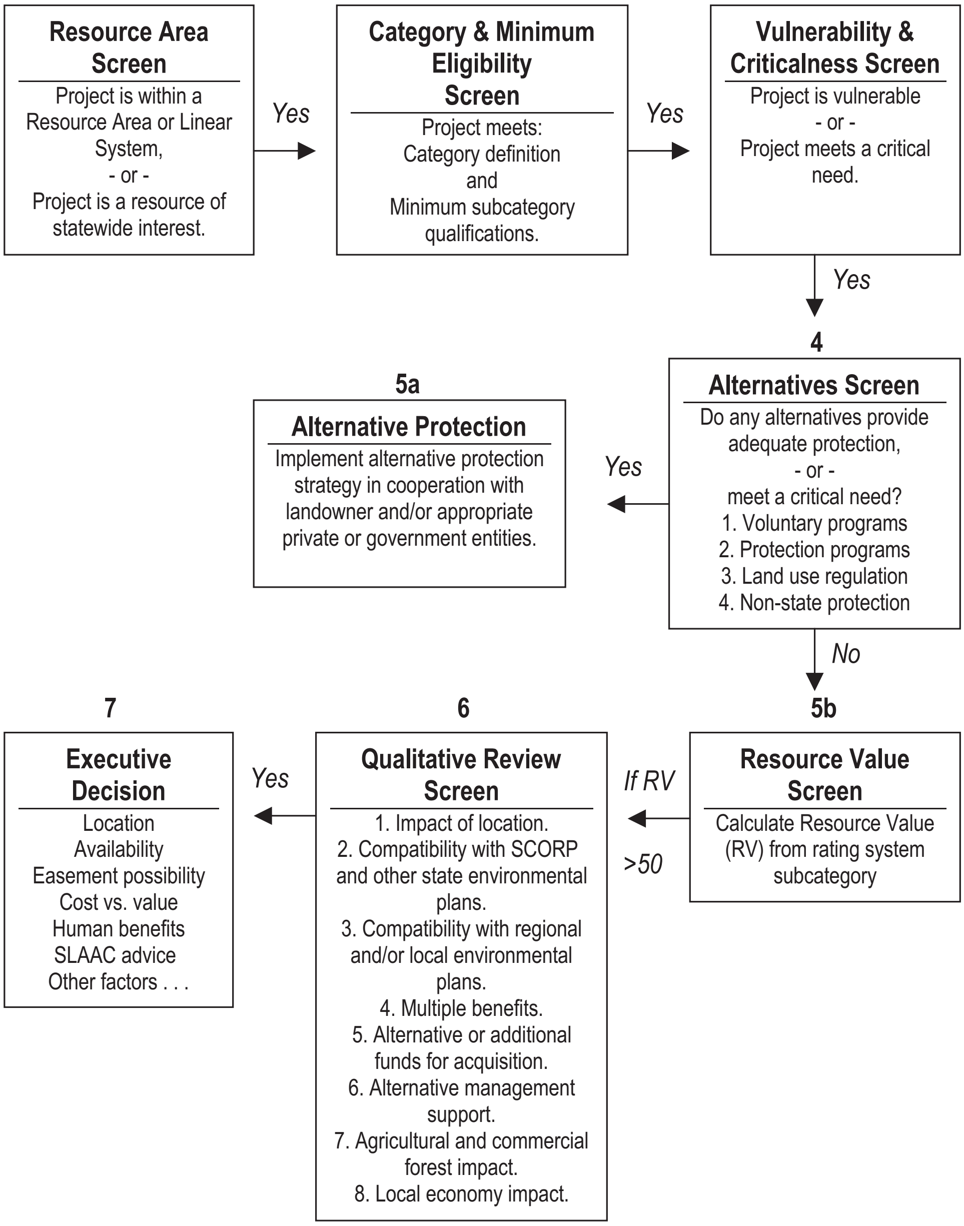


Figure 6.1. Project review and selection process.
Notes: SLAAC = State Land Acquisition Advisory Council; SCORP = Statewide Comprehensive Outdoor Recreation Plan.
Source: DEC and OPRHP, 1998, Figure 13.

most recent State Open Space Conservation Plan. The Adirondack Park is one of nine major resource areas identified in the 1998 Plan. Linear systems include areas that extend along continuous natural features (i.e., rivers or mountain ridges) or trail corridors. Examples include the Hudson River Valley, the New York State Canal Recreationway System, and the Appalachian Trail (a portion of which crosses New York State on its east coast journey from Georgia to Maine). Even if a proposed project does not fall within one of these predefined conservation targets, it can pass this screen if the parcel is considered to be a resource of statewide interest.

Next, a proposed parcel must fall within a conservation category and meet minimum subcategory qualifications. *Table 6.1* lists the six protection categories along with the 21 specific subcategories. The major categories are forest preserve addition, water resources protection, significant ecological area, recreational opportunity, distinctive character, and enhancement of public lands. Under the current system, a parcel can only be categorized under one subcategory, even though it may have attributes that qualify it for many. Minimum requirements differ widely amongst subcategories. A forest preserve addition can only be considered if it is located within either the Adirondack or Catskill Park and outside the boundaries of an incorporated village or city. A forest and scenic easement must protect productive forestland located within either the Adirondack or Catskill Park. Requirements under other sub-criteria tend to be much more specific than these two forest-preserve categories.

The third stage in the screening process ascertains the degree of urgency for protection, considering for instance the present condition of the site, any pending ownership transfer, the relationship with any local land use plans, and the land use pattern and development trends in the area. In addition, consideration is given to the compatibility of a proposed parcel with objectives other than preservation (i.e., access, resource management) and the availability of alternative sites to meet those objectives.

Once a parcel passes through these initial screens, the regional DEC office must then determine whether alternatives to state purchase or easements exist that can still provide adequate protection or meet a critical need. For instance, voluntary private conservation or enrollment in non-state protection programs may satisfy a particular objective. In recent years, the State has relied heavily on groups such as the Nature Conservancy to protect key parcels because either acquisition funds or ability to acquire a parcel in a timely manner is lacking. The role of land trust organizations in pre-acquisition of permanent State property accounted for 22% of transactions eventually acquired under 1986 Bond Act funds, amounting to 79% of acreage purchased and 68% of dollars spent (The Land Trust Exchange and Russel, 1990, pp. 172–186). The state may also consider regulation versus outright purchase to protect a parcel from development or unsustainable use. If a feasible

Table 6.1. Land protection categories and sub-categories.

Forest Preserve	Water Resource Protection	Significant Ecological Area	Recreational Opportunity	Distinctive Character	Enhancement of public lands
Forest preserve additions	Aquifer recharge area	Exceptional forest	Metropolitan parks and shorelines	Historic preservation	Access
Forest land easements	Watershed protection	Shoreline protection	Parklands	Working landscapes	Buffer
		Unique areas	Public fishing rights	Heritage areas	Consolidation and inholdings
		Wetlands	Trailways and and greenways	Scenic resources	
		Wildlife habitat	Waterway access		

Source: DEC and OPRHP, 1998, Table XI.

alternative can be negotiated without the use of state acquisition funds, then this provides an opportunity to exit the evaluation process outlined in *Figure 6.1*.

At this point in the process, if State acquisition of land or development rights seems like the best course of action then the parcel under consideration enters a formal resource value screen (Stage 5b). The rating is a numerical score, unique to each subcategory, assigned by professional staff (typically a DEC forester) on a scale of 0 to 100 points. A minimum of 50 points is required for consideration under the Open Space Conservation program. Again, a parcel can only be evaluated applying one of the twenty-one subcategories outlined in *Table 6.1*. However, any project meeting the minimum criteria of one other subcategory receives an additional three points. If more than one additional subcategory applies, then five points are awarded. In addition, gifts of land avoid purchase costs and are thus awarded 10 extra points towards the 100 total. *Table 6.2* outlines the resource value-rating scheme for the subcategory of forest preserve easements.

The point system is not meant to compare parcels between different categories or subcategories. Resource values are only comparable within unique subcategories. Under the present system, subcategory scores can be used to rank acquisition, but mainly serve as a threshold before projects are recommended for a final screening. If a project receives a resource value score of at least 50 points then it is eligible to move into the qualitative review screen.

Table 6.2. Resource value rating system: Adirondack and Catskill Park forest and scenic easements.

Characteristic	Rating
a. Proposed project will provide new or enhance existing recreational opportunities.	
(i) Choose one:	
(a)Project provides five or more opportunities for a variety of both land and water related recreational activities;	10
(b)project provides between two and five opportunities for a variety of either land or water related recreational activities; or	5
(c)project provides for a single purpose recreational opportunity of either a land or water related activity.	1
(ii) Choose one:	
(a)Project provides alternate recreational opportunities for an existing recreational area which is currently experiencing high use;	10
(b)project provides recreational opportunity to a geographical area where there is a demand for recreational use but which currently has little or no recreational opportunity; or	10
(c)project provides additional opportunity to an area which is not presently experiencing high use.	1
b. The proposed project's maximum value is:	(30)
(i)protects threatened or endangered plant or animal species	10
(ii)protects significant habitats	10
(iii)protects rare natural communities	10
(iv)protects Class I regulated wetlands; or	10
(v)protects undeveloped shorelines of importance. Importance is defined by designation as: 1. a wild, scenic or recreational river; 2. critical environmental area; 3. scenic area of statewide importance; or 4. national natural landmark.	10
c. Proposed project protects recognized scenic areas or views, including scenic highway corridors that require the manipulation of vegetation to preserve.	5
d. Proposed project provides or enhances access to inaccessible or poorly accessible portions of Forest Preserve or other lands or waters.	
(i)the proposed project would provide access or assist in providing access to public lands or waters which presently have no existing access open to the public; or	5
(ii)the proposed project would provide access or assist in providing access to public lands or waters to which existing access is poor because of physical barriers; or	3
(iii)the proposed project would reduce the length of a circuitous route of three miles or more necessary for public use of existing public lands or waters.	1

Table 6.2. Continued.

Characteristic	Rating
e. The value of the continuation of forestry uses is determined by application of the following rating scale. The maximum value is:	(40)
(i) productivity factor: rate the overall productivity of the project using such factors as soils, income potential, species composition, products produced, significance to industry, and other relevant factors:	
(a) high,	20
(b) medium,	10
(c) low.	5
(ii) survival factor: rate the likelihood of the project continuing in present use using such factors as: capital investment, product demand, owner commitment, accessibility, and other relevant factors:	
(a) high,	20
(b) medium,	10
(c) low.	5
f. The present degree of development and extent of viewshed proposed for protection is determined by the application of the following rating scale. The maximum value is:	(40)
(i) current degree of development as expressed as a percent of maximum buildout allowed under existing zoning:	
(a) <= 20%	20
(b) >20% and <=50%	10
(c) >50% and <=70%	5
(ii) ratio of project acreage within either 500 feet of mean high water or 1,000 feet of public viewing point (highway, trail, etc.) to total project acreage is not less than 40%:	20
(a) >75%	10
(b) 60% to <75%	
(c) >=40% and <60%	5

Source: DEC and OPRHP, 1998, Appendix C.

Eight criteria are used at this final stage to justify a formal acquisition or easement proposal to the Commissioner of the DEC. The first six criteria are similar to considerations taken in screens one through four. At this point, considerations of project compatibility, multiple benefits, and the fund source and mechanics of title acquisition are made more explicit. However, the seventh and eighth criteria explicitly consider economic impacts of parcel acquisition for the first time in the review and selection process. At this stage, staff of either the DEC or OPRHP follow a checklist that was developed to help evaluate potential fiscal and economic benefits and burdens associated with a proposed project (DEC and OPRHP, 1998, p. 66). These factors include the project's impact on: real property tax base; local and regional retail sales and service businesses; real estate values; traffic flow; land use patterns; funding by bonding, direct allocation, gift, federal funds, or private funding sources; and farming and forestry resource base in the town or county.

The ultimate recommendation to the Commissioner follows careful consideration of data from each of the six screens, comment from local government, and any input from the State Land Acquisition Advisory Council (SLAAC). The Commissioner decides whether or not to proceed with acquisitions and has the discretion to rank approved projects.

If the current procedure is taken at face value (ignoring political realities of a very flexible process for the moment), there are a number of shortcomings that can be identified. First, the shortcoming of being able to compare only land parcels within the same sub-category could be overcome, if a sound multi-criteria analysis were the basis of ranking the parcels. Possible incompatibilities between categories can still be taken into consideration under such a framework. Second, the evaluation of the criteria for the land parcels seems rather ad hoc and subjective instead of being based on sound scientific information. Third, there seems to be some degree of misplaced concreteness involved in the evaluation of the criteria. For example, it seems difficult to argue exact differences for the criteria "scenic resources." Aesthetics are usually best expressed in linguistic variables that are, however, best translated into fuzzy variables instead of crisp ones. Fourth, other criteria could be included which would probably increase the acceptability of the evaluation scheme within the population in the area. Such criteria could include economic variables such as estimates of resulting job creation/destruction effects stemming from land use changes, or social criteria like residential attractiveness. Fifth, the transparency of the decision making process should be increased, i.e. the criteria and their evaluation laid open to the public. Sixth, particularly with a history of struggle and dispute as is the case in the Adirondack Park, the involvement of all relevant stakeholders is crucial for achieving widely accepted solutions. Stakeholder input currently is only included before the formal project screening occurs (i.e., in the pre-screening of projects by regional open space committees who recommend formal evaluation).

6.3 Alternative Problem Structuring with Multicriteria Decision Aid

Decision making on sustainable land use usually involves competing interest groups, conflicting objectives, and different types of information. Multi-criteria decision aid (MCDA) is a tool that can be used to consider simultaneously multiple conflicting criteria (e.g., representing economic, environmental, social, institutional, technical, and aesthetic objectives). The aim is "to enable us to enhance the degree of conformity and coherence between the evolution of the decision-making process and the value systems and objectives of those involved in this process" (Roy, 1990, p. 17). This points to the importance of the decision makers in this process, but also the fact that the result of an MCDA method is only an input into the decision-making process and not the final result.

6.3.1 NAIADE algorithm and software

The multidimensionality is also a characteristic of the scenario of open space acquisition under investigation. For this reason MCDA is used. Specifically, the NAIADE (Novel Approach to Imprecise Assessment and Decision Environments) method (developed by Munda, 1995) was found to be effective in this specific case for several reasons.

NAIADE belongs to the group of discrete multicriteria methods, i.e., the set of alternatives is finite (for a good overview of methods see Vincke, 1992). Using a pairwise comparison technique, NAIADE generates a ranking of alternatives according to the set of evaluation criteria. The comparison of criteria scores of each pair of alternatives is carried out by means of semantic distance which mirrors a possible degree of equality between two fuzzy sets or a similarity degree between them; the larger the distance the smaller the possible degree of equality. Fuzzy binary relations are used to model different possible preference/indifference situations. The aggregation of the evaluations of the alternatives according to each single criterion is done such that the intensity of preference is incorporated.

More specifically, the intensity index $\mu_*(a,b)$ of preference $*$ (where $*$ stands for $>>$, $>$, $\cong$, =, $<$ and $<<$) of alternative a versus b is defined as follows (Munda, 1995:137n.):

$$\mu_*(a,b) = \frac{\sum_{m-1}^{M} \max(\mu_*(a,b)_m - \alpha, 0)}{\sum_{m-1}^{M} \mid \mu_*(a,b)_m - \alpha \mid} \ .$$

The intensity index $\mu_*(a,b)$ has the following characteristics:
$0 \leq \mu_*(a,b) \leq 1$
$\mu_*(a,b) = 0$ if none of the $\mu_*(a,b)_m$ is greater than α;
$\mu_*(a,b) = 1$ if $\mu_*(a,b)_m \geq \alpha \forall m$, and $\mu_*(a,b)_m \geq \alpha$ for at least one m.

The parameter α, which can be changed in the analysis, is the 'minimum requirement' imposed on the fuzzy relation to distinguish between different degrees of preference and indifference in the aggregation (Munda, 1995). This means that with increasing α only values having a high intensity of preference or indifference are used. Or more precisely, only those criteria whose indexes are above the threshold will be counted positively in the aggregation (Menegolo and Pereira, 1996).

Moreover, when α increases, a lower degree of compensation among the criteria is allowed. If too high or too low values are used, it is difficult to discriminate between actions (Munda, 1995).

The ranking of alternatives in NAIADE is based on the preference intensity indexes $\mu_*(a, b)$ and corresponding entropies $H_*(a, b)$ for the alternatives a and b. The ranking process is based on the basic idea of positive (leaving) and negative (entering) flows of the PROMETHEE methods (Brans *et al*., 1986). A partial ranking of alternatives can be deduced from the positive (ϕ^+) and the negative (ϕ^-) outranking flows (see PROMETHEE I). Both rankings are usually not identical. The final ranking comes from the intersection of the two separate rankings. The first one $\phi^+(a)$ is based on the better and much better preference relations; its value ranges from 0 to 1 indicating how a is better than all other alternatives. The second outranking flow, $\phi^-(a)$, is based on the worse and much worse preference relations; its value ranges from 0 to 1 indicating how a is worse than all other alternatives (Menegolo and Pereira, 1996).

In comparison, the much more widely used method of 'Analytic Hierarchy Process' (AHP) is based on the construction of hierarchies and pairwise comparisons that are used for establishing weights. Since AHP is based on measuring preferences cardinally, its underlying ideas differ significantly from the ones of NAIADE. Also, AHP does not address uncertainty.

The NAIADE method is a recently developed MCDA approach, whose impact matrix can include crisp, stochastic, or fuzzy measurements of the performance of each option with respect to a judgment criterion. No weighting of criteria is used explicitly (Munda, 1995). Hence, it allows the use of information affected by different types of uncertainty. In addition to the ranking of alternatives, NAIADE supports the analysis of conflicts between different interest groups and the possible formation of coalitions according to the proposed alternatives. The method is implemented by a software application also called NAIADE (for case studies applying this method see, for example, De Marchi *et al*., 2000; De Montis *et al*. 2000).

The NAIADE method is used in this case study for several reasons. First, the current evaluation procedure consists of several steps that are based on different types of information. The impact (or evaluation) matrix in NAIADE may include either crisp, stochastic, or fuzzy measurements of the performance of each option with respect to a judgment criterion (Munda, 1995). Some of the criteria (like acquisition costs, loss of agricultural land, or impact on retail sales) can be measured in quantitative terms. Others (like protection of scenic area or multiple benefits) are expressed in qualitative terms. In order to incorporate this diverse information, a method was necessary that incorporates both types of data. In addition, the information may be available – as it is in the cases under consideration – in rather rough categories. While unsatisfactory from a scientific point of view, the data may not be available in a more precise way (or too expensive to be gathered). To include rough categories into a transparent and consistent analysis is preferable to dropping the information completely or to including unfounded information. In other cases, it may be impossible to express criteria in concrete numbers absent of fundamental uncertainty. In particular, criteria on the interface between the social and the environmental system may be greatly affected by uncertainty.

The second reason for the choice of NAIADE over other multi-criteria techniques is the ability to conduct conflict analysis. In addition to the ranking of alternatives, NAIADE supports the analysis of conflicts between different interest groups and the possible formation of coalitions according to the proposed alternatives. This may help to make the decision process more transparent, and will be explored in a future extension of this work.

Furthermore, the selection of operators and choice of parameters allows us to apply the software to problems where differing degrees of compensation of criteria performance is desired and to test for sensitivity of the results.

6.3.2 Data

The criteria were given in the *Open Space Plan*. Using the evaluations of five recently considered parcels from the sub-category 'forest easement,' an impact matrix was constructed. The data were provided by the DEC of New York State and complemented wherever necessary by expert opinion.

As can be seen in *Table 6.3*, all variables were defined as linguistic variables. The data from the Qualitative Review Screen were only available in this way. Even the points assigned in the Resource Value Screen are mere representations of a discrete number of linguistic evaluations. This view is supported by the fact that points are only assigned in discrete steps and not on a continuous scale of numbers. To assure that the decision-makers' preferences are accounted for, we kept the distances between and the different weights of the respective points (for details see notes to *Table 6.3*). In NAIADE, the linguistic variables are defined by means of fuzzy sets defined by a 0 to 1 scale, whereby 1 indicates 'perfect' and 0 indicates 'extremely bad.'

Unfortunately, the data describing parcels that were not eventually acquired in fee or easement is not archived by the DEC. Therefore it is not possible to analyze a complete decision situation and to compare the administrative decision with the results of the model-based decision framework.

If more precise information were available for some of the criteria – costs, for example – (see Notes to *Table 6.3*) it could be introduced into NAIADE in real numbers or ranges of numbers (fuzzy sets).

Table 6.3. Evaluations for five parcels with the criteria from Resource Value Screen and Qualitative Review Screen.

	Criteria	Long Pond Tract (A)	Otetiana (B)	Santa Clara Tract (C)	Tooley Pond Tract (D)	Croghan Tract (E)
1.	Types of recreation	more than 5	more than 5	more than 5	more than 5	more than 5
2.	Complementarity to existing recreation opportunities	w/o high use	w/o high use	w/o high use	no	high use
3.	Protection endangered species	no	yes	no	no	no
4.	Protection significant habitat	yes	yes	no	no	yes
5.	Protection rare natural communities	yes	yes	yes	yes	yes
6.	Protection wetlands	no	no	no	no	no
7.	Protection shorelines	no	no	no	no	no
8.	Protection scenic area	no	yes	no	no	no
9.	Improvement of accessibility	little	no	no	no	little
10.	Productivity factor	high	medium	high	high	high
11.	Survival factor	high	medium	high	high	high
12.	Impact of land use patterns	weak no	no	no	weak no	no
13.	Conflicts w/other State plans	no	no	no	no	no
14.	Conflicts w/environmental plans	no	no	no	no	no
15.	Multiple benefits	weak yes	yes	weak yes	weak yes	weak yes
16.	Alternative/additional funding sources	no	potentially	no	no	no
17.	One time costs	low	low	high	high	high
18.	Future annual costs	low	low	high	high	high
19.	Possibility to share costs	no	potentially	potentially	potentially	potentially
20.	Agricultural land loss	no	no	no	no	no
21.	Impact on local tax base	weakly positive	positive	very positive	very positive	very positive
22.	State paying real property tax	yes	partially	partially	partially	partially
23.	Impact on retail sales/service business	neutral	somewhat positive	somewhat positive	weakly positive	weakly positive
24.	Impact on local real estate values	neutral	weakly positive	neutral	neutral	neutral

Table 6.3. Continued.

	Criteria	Long Pond Tract (A)	Otetiana (B)	Santa Clara Tract (C)	Tooley Pond Tract (D)	Croghan Tract (E)
25.	Impact on traffic flow	neutral	weakly negative	weakly negative	neutral	neutral
26.	Impact on local land use patterns	no	no	no	no	no
27.	Direct cost to NYS tax payer	negative	negative	negative	negative	negative
28.	Direct cost to local tax payer	positive	positive	positive	positive	positive
29.	Impact on farming/resource base	positive	neutral	positive	positive	positive

Notes: To account for preference intensity equivalents, points were assigned to the nine-part scale of qualitative evaluations suggested by the software. Hence, the highest value represents 20 points in the Resource Value Screen, decreasing at equal distances to zero (relevant for criteria 1 to 11). The only difficulty with this procedure arose for values 1 and 3 where 2.5 had to be assigned as an approximation. Criteria 12 to 29 came from the Qualitative Review Screen. The options for the questions related to criteria 12 and 15 are: 'absolutely,' 'yes,' 'weak yes,' 'maybe,' 'not certain,' 'don't think so,' 'weak no,' 'no,' 'no way.' The options for the questions related to criteria 13, 14, 16, 19 and 20 are: 'yes,' 'potentially,' 'no.' The options for the questions related to criteria 17 and 18 are: 'low,' 'medium,' and 'high.' The options for the questions related to criteria 21, 23, 24 and 25 were 'very positive,' 'positive,' 'somewhat positive,' 'weakly positive,' 'neutral,' 'weakly negative,' 'somewhat negative,' 'negative,' and 'very negative.' The options for the questions related to criteria 27, 28 and 29 are: 'positive,' 'neutral,' and 'negative.' The options for the questions related to criteria 22 and 26 are: 'yes,' 'partially,' and 'no.' All criteria are maximized except 'one-time costs' (17), 'future annual costs' (18) and 'agricultural land loss' (20), which are minimized.

6.3.3 Results and discussion

A critical factor in determining the results provided by the NAIADE method is the parameter used in the equation on approximate reasoning operations.

By use of the minimum operator, which is known as a representation of the logic 'and,' the ranking obtained for a low value of α (0.3) is given in *Figure 6.2*. The NAIADE program computes separate rankings for the positive and negative outranking flows with their respective values. The higher the value of the positive outranking flow, the higher its 'power,' i.e., the better one alternative is compared to the others. In our case, A (project: Long Pond Tract) is better than B (project: Otetiana) which is better than E (project: Croghan Tract), etc. The higher the value of the negative outranking flow the higher its 'weakness', i.e., the worse is one alternative compared to the others. Here, B is worse than C (project: Santa Clara Tract) which is worse than D (project: Tooley Pond Tract), etc. Hence, the higher the value of $\phi^{+}(a)$ and the lower $\phi^{-}(a)$, the better is alternative a. The final ranking

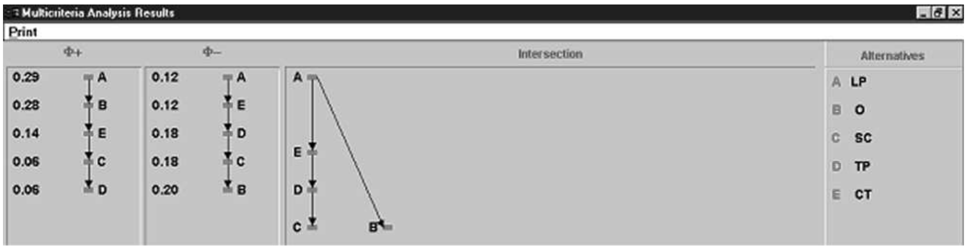


Figure 6.2. Ranking of parcels (minimum operator, $\alpha = 0.3$).

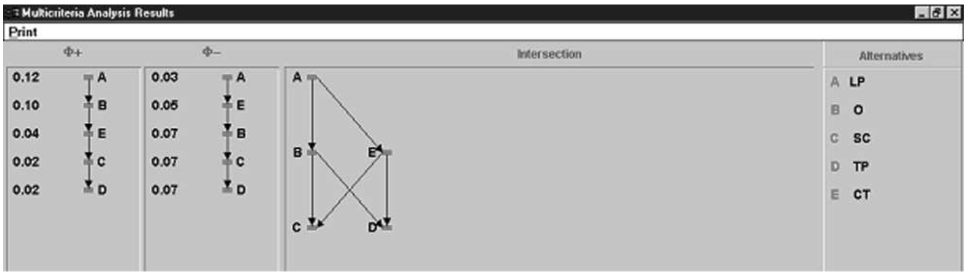


Figure 6.3. Ranking of parcels (minimum operator, $\alpha = 0.5$).

is obtained from the intersection of the two outranking flows. A has the highest positive outranking flow and the lowest negative outranking flow and is therefore preferred to all other alternatives. B has a high positive outranking flow but also a high negative outranking flow, i.e. it is on the one side better than the three other alternatives, but also worse then these alternatives, and therefore not comparable. The difference between alternatives A and E in the negative outranking flow is very small (less than 0.01). However, A is significantly better than E in the positive outranking flow. In total, E is therefore dominated by A. Differences in the values of both outranking flows of alternatives C and D are very small. The domination of D over C is therefore very weak.

Increasing the value of α (0.5) (see *Figure 6.3)* increases incomparabilities, but the main findings remain the same. A has a higher positive outranking flow and a lower negative outranking flow than the other alternatives and is therefore preferred to the others. While B has a higher positive outranking flow than E, alternative E has a lower negative outranking flow. These two alternatives are therefore incomparable. Alternatives C and D are dominated by all the other alternatives. Differences between them are too small, therefore they are also incomparable in this specification.

Since the outranking flows are already quite low, increasing α further is not recommendable.

The pairwise linguistic evaluations give indications of the relative credibility degree of preferences and therefore complements the ranking which is ordinal in nature. The alternatives considered here were all successful ones and therefore the evaluations are very similar, hence it is not surprising that the differences between most parcels seen through the pairwise comparisons are not very high.

In sum, it can be seen that NAIADE is a tool that can help decision-making in this case by providing rankings allowing for different degrees of compensation between the values of the fuzzy relations. The results vary to some degree with the specifications, but not in the main findings. The selection of specifications reasonable in this context needs to be done by the decision-makers.

The framework and procedure presented here allow the inclusion of other features that may be useful for better decision making. First, if more precise information were available, this could be included either as real numbers or at least as ranges of values (fuzzy numbers). Second, the application of similar criteria to all sub-categories would enable decision-makers to compare parcels across categories in a coherent and (publicly) defendable way. The criteria, however, need to be global, i.e., applicable to all categories, because incomparabilities will result otherwise. This does not mean that no distinction could be made between the characteristics of the different sub-categories. The criteria would have to be defined broadly enough and could then be filled with the information adequate for the respective sub-category.

On a different level, the decision process could be improved by integrating different groups of stakeholders into the decision-making process. Besides an "impact matrix," each group also constructs an "equity matrix," which contains linguistic evaluations of alternatives. In particular, "equity analysis is performed by the completion of an equity matrix from which a similarity matrix is calculated. Through a mathematical reduction algorithm, it is possible to build a dendrogram of coalitions which shows possible coalition formation, and a level of conflict among the interest groups" (Menegolo and Pereira, 1996, p. 1).

Unfortunately, the information necessary to do such an analysis was not available in our case. The inclusion of stakeholders in a transparent process could, however, increase the acceptability and defendability of the decision.

6.4 Conclusions

Due to major economic structural changes, large tracts of private land are currently for sale in the Adirondacks. The state authorities used land acquisition and conservation easements, among other instruments, to enhance sustainable land use in the Adirondack State Park. The goals, the extent, and the process have often been criticized for a lack of transparency and consistency and resulted in fierce disputes among the various interest groups. In order to address these criticisms this chapter suggests a framework that applies MCDA.

MCDA helps to structure the decision-making process and the relevant information. It increases the transparency of the process and provides an algorithm for ranking parcels. Its ability to include quantitative and qualitative criteria within a consistent framework is particularly useful. Even where quantitative data exist they are very often qualitative in nature and should and can be treated as such.

NAIADE has proven to be particularly suitable in this context because the uncertainty inherent in sustainability questions is addressed with the concept of fuzzy sets as used in the evaluation matrix. Despite the imprecise information, NAIADE allows a consistent evaluation without imposing strong assumptions. The structure of the method shows weaknesses in data and shows the direction of further data collection. Furthermore, the diverse values expressed by different stakeholders can be included with the addition of an equity matrix, which can highlight coalition potential in conflicting situations. The information provided by the stakeholders and the analysis of their positions is a valuable input into the process towards an acceptable decision. Coalitions and values are made explicit and therefore allow an open discussion of assumptions and valuations.

The current analysis has been restricted due to the unavailability of data on parcels not chosen after the review processes. Inclusion of those parcels would enable transparency of the current decision-making process and would increase the acceptability by all involved or affected. This paper has relied on documented information.

A discussion with decision-makers and stakeholders is a necessary next step to discuss questions of desired compensability between criteria. A difficult issue remains to be explored in the application: How deep an insight do users need to get into the sophisticated technicalities of the method (e.g., the ranking procedure or the concept of fuzzy sets) in order to feel comfortable using results from the analysis? It is our conviction that the appropriate choice of technical specifications, their translation by the researcher into non-technical language, and the discussion of crucial specifications is essential for acceptable results. It is the responsibility of the researcher to ensure this through non-technical discussions with stakeholders. Hence, we do not find that simple methods which require strong and unrealistic assumptions should be preferred. In the past, NAIADE has been applied successfully in several cases with stakeholder involvement (e.g., De Marchi *et al*., 2000; Race, 2000).

This chapter represents a valuable contribution in the evaluation process and provides a starting point for reevaluating the decision-making process as well as a procedure to include the groups concerned.

References

Brans, J.-P., Mareschal, B., and Vincke, P., 1986, How to select and how to rank projects: The PROMETHEE method, *European Journal of Operational Research*, **24**:228–38.

DEC (Department of Environmental Conservation) and OPRHP (The Office of Parks, Recreation and Historic Preservation), 1998, Conserving Open Space in New York State, Open Space Conservation Plan and Final Generic Environmental Impact Statement, Albany, NY, USA.

De Marchi, B.S.O., Funtowicz, S.O., Lo Cascio, S., and Munda, G., 2000, Combining participative and institutional approaches with multicriteria evaluation: An empirical study for water issues in Troina, Sicily, *Ecological Economics*, **34**:267–82.

De Montis, A., De Toro, P., Droste-Franke, B., Omann, I., and Stagl, S., 2000, Criteria for quality assessment of MCDA methods, 3rd Biennial Conference of the European Society for Ecological Economics, Vienna University of Economics and Business Administration, Vienna, Austria.

Erickson, J.D., 1998, Sustainable development and the Adirondack Park experience, *Adirondack Journal of Environmental Studies*, **5**(2):24–32.

Erickson, J.D., and O'Hara, S.U., 2000, From top-down to participatory planning: Conservation lessons from the Adirondack Park, NY, USA, in L. Tacconi, ed., *Biodiversity and People: An Ecological Economic and Participatory Approach*, Earthscan, London, UK.

Menegolo, L., and Pereira, A., 1996, *NAIADE: Manual*, Joint Research Centre of the European Commission, Ispra, Italy.

Munda, G., 1995, *Multicriteria Evaluation in a Fuzzy Environment: Theory and Applications in Ecological Economics*, Physika Verlag, Heidelberg, Germany.

Race, S., 2000, *Strumenti e valutazioni per lo sviluppo sostenibile del Fondo Cinque Lenze e dei territori afferenti al Lago Miseno in Bacoli*, Dissertation at Federico II University of Naples, Faculty of Architecture, Naples, Italy.

Roy, B., 1990, Decision-aid and decision-making, in C.A. Bana e Costa, ed., *Readings in Multiple Criteria Decision Aid*, Springer Verlag, New York, pp. 17–35.

The Land Trust Exchange and Joel S. Russel (Woodlea Associates), 1990, *The Role of Nonprofit Third Parties in State Land Acquisition*, Technical Report 8, Commission on the Adirondacks in the Twenty-First Century, Vol. 1, Appendix, State of New York, USA.

Vincke, P., 1992, *Multicriteria Decision-Aid*, Wiley and Sons, Chichester, UK, and New York, NY, USA.

Part II: Data Analysis

Chapter 7

Rough Sets and Intelligent Data Analysis

Zdzisław Pawlak

Abstract

Rough-set-based data analysis starts from a data table, called an *information system*. The information system contains data about objects of interest characterized in terms of some attributes. In the information system we often distinguish condition and decision attributes. Such an information collection is called a *decision table*. The decision table describes decisions in terms of conditions that must be satisfied in order to carry out the decision specified in the decision table. Associated with every decision table is a set of decision rules, called a *decision algorithm*. Every decision algorithm reveals some well known probabilistic properties; in particular, it satisfies the Total Probability Theorem and the Bayes' Theorem. These properties give a new method of drawing conclusions from data, without referring to prior and posterior probabilities, inherently associated with Bayesian reasoning.

Keywords: Vagueness, uncertainty, rough sets, data mining, data analysis, environmental management, global warming.

7.1 Introduction

Rough set theory is a new mathematical approach to intelligent data analysis and data mining. After almost 20 years of pursing rough set theory and its application, the approach reached a certain degree of maturity. In recent years we have witnessed a rapid growth of interest in rough set theory and its application, worldwide. Many international workshops, conferences, and seminars have included rough sets in their programs. So far, about 2,000 papers and several books have been published on various aspects of rough sets.

Rough set philosophy is founded on the assumption that we associate some information (data, knowledge) with every object in the universe of discourse. Objects characterized by the same information are *indiscernible* (*similar*) in view of the available information about them. The *indiscernibility relation* generated in this way is the mathematical basis of rough set theory. Any set of indiscernible (similar) objects is called an *elementary set*, and forms a basic *granule* (*atom*) of knowledge about the universe. Any union of elementary sets is referred to as a *crisp* (*precise*) set; otherwise, the set is *rough* (*imprecise*, *vague*). Each rough set has boundary-line cases, i.e., objects that cannot be classified with certainty by employing the available knowledge, either as members of the set or its complement. Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. A pair of precise sets – called the *lower* and the *upper approximation* of the rough set – is associated with any rough set. The lower approximation consists of all objects which *surely* belong to the set, and the upper approximation contains all objects which *possibly* belong to the set. The difference between the upper and the lower approximation constitutes the *boundary region* of the rough set. Approximations are two basic operations in rough set theory.

The rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning, and pattern recognition. Rough set theory has been successfully applied in many real-life problems in medicine, pharmacology, engineering, banking, finances, market analysis, environmental management, and others.

The rough set approach to data analysis has many important advantages. It:

- provides efficient algorithms for finding hidden patterns in data,
- finds minimal sets of data (data reduction),
- evaluates the significance of data,
- generates sets of decision rules from data,
- offers straightforward interpretation of obtained results,
- yields algorithms that are particularly suited for parallel processing, and
- is easy to understand.

The basic concept of rough-set-based data analysis will be outlined below and will be illustrated by a simple tutorial example concerning global warming.

The application of rough sets to environmental management can be found in Flinkman *et al*. (to appear), Grzymała–Busse and Grzymała–Busse (1994), Grzymała–Busse (1994), Grzymała–Busse and Gunn (1995), Gunn and Grzymała–Busse (1994), Keiser *et al*. (1992), Reinhard *et al*. (1992), Teghem and Charlet (1992), and Zhong *et al*. (1999). The basics of rough sets can be found in Düntsch and Günter (2000) and Pawlak (1991).

More about rough sets can be found in the references and on the Web. A list of relevant sites appears at the end of this chapter.

7.2 Information Systems and Decision Tables

The starting point of rough-set-based data analysis is a data set, called an information system. An information system is a data table: its columns are labeled by attributes, its rows are labeled by objects of interest, and its entries are attribute values.

Formally, by an *information system* we will understand a pair $S = (U, A)$, where U and A are finite, nonempty sets called the *universe*, and the set of *attributes*, respectively. With every attribute $a \in A$ we associate a set V_a, of its *values*, called the *domain* of a. Any subset B of A determines a binary relation $I(B)$ on U, which will be called an *indiscernibility relation*, and defined as follows: $(x, y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in B$, where $a(x)$ denotes the value of attribute a for element x. Obviously, $I(B)$ is an equivalence relation. The family of all equivalence classes of $I(B)$, i.e., a partition determined by B, will be denoted by $U/I(B)$, or simply by U/B; an equivalence class of $I(B)$, i.e., block of the partition U/B, containing x will be denoted by $B(x)$.

If (x, y) belongs to $I(B)$ we will say that x and y are *B-indiscernible* (*indiscernible with respect to B*). Equivalence classes of the relation $I(B)$ (or blocks of the partition U/B) are referred to as *B-elementary sets* or *B-granules*.

If we distinguish in an information system two disjoint classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision table* and will be denoted by $S = (U, C, D)$, where C and D are disjoint sets of condition and decision attributes, respectively.

An example of a decision table is shown in *Table 7.1*, where *Solar Energy*, *Volcanic Activity* and *Residual* CO_2 are condition attributes, and *Temperature* is a decision attribute.

The example concerns global warming and is taken, after some simplifications, from Grzymała–Busse (1994).

Table 7.1. An example of a decision table.

Fact	Solar energy	Volcanic activity	Residual CO_2	Temperature	Days count
1	medium	high	low	high	20
2	high	high	high	high	30
3	medium	low	high	high	90
4	low	high	low	low	120
5	high	low	medium	high	70
6	medium	low	high	low	34

Source: Modified from Gryzmała–Busse (1994).

We want to explain what causes the high (low) temperatures, i.e., to describe the set of facts {1, 2, 3, 5} ({4, 6}) in terms of condition attributes: *Solar Energy, Volcanic Activity and Residual* CO_2.

The data set is *inconsistent* because facts 3 and 6 are contradictory, therefore the problem cannot be solved exactly but only approximately.

Let us observe what the data are telling us:

- Facts 1, 2, 5 can be *certainly* classified as causing high temperature.
- Fact 4 can be *certainly* classified as causing low temperature.
- Facts 3, 6 can be *possibly* classified as causing high or low temperature.

7.3 Approximation of Sets

Suppose we are given an information system $S = (U, A)$, $X \subseteq U$, and $B \subseteq A$. Our task is to describe the set X in terms of attribute values from B. To this end we define two operations assigning to every $X \subseteq U$ two sets $B_*(X)$ and $B^*(X)$ called the *B-lower* and the *B-upper approximation* of X, respectively, and defined as follows:

$$B_*(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\} \; , \tag{7.1}$$

$$B^*(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \emptyset\} \; . \tag{7.2}$$

Hence, the *B*-lower approximation of a set is the union of all *B*-granules that are included in the set, whereas the *B-upper* approximation of a set is the union of all *B*-granules that have a nonempty intersection with the set. The set

$$BN_B(X) = B^*(X) - B_*(X) \tag{7.3}$$

will be referred to as the *B-boundary region* of X.

If the boundary region of X is the empty set, i.e., $BN_B(X) = \emptyset$, then X is *crisp* (*exact*) with respect to B; in the opposite case, i.e., if $BN_B(X) \neq \emptyset$, X is referred to as *rough* (*inexact*) with respect to B.

For our illustrative example we have, with respect to the condition attributes:

- The set {1, 2, 5} is the *lower approximation* of the set {1, 2, 3, 5}.
- The set {1, 2, 3, 5, 6} is the *upper approximation* of the set {1, 2, 3, 5}.
- The set {3, 6} is the *boundary region* of the set {1, 2, 3, 5}.

The above considerations are illustrated in *Figure 7.1*.

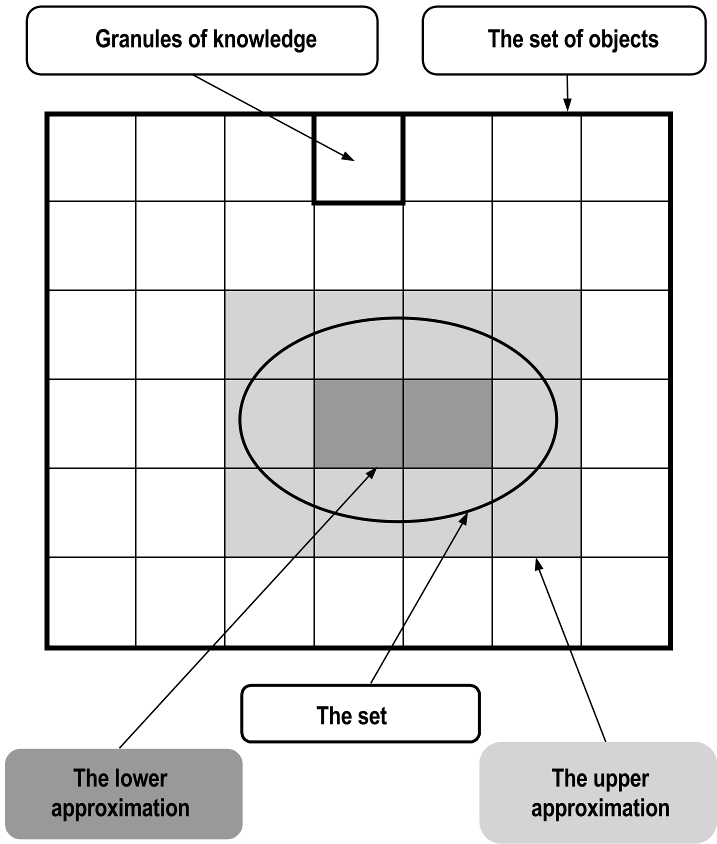


Figure 7.1. Approximation of sets.

7.4 Decision Rules

Decision rules constitute a formal language for describing approximations in logical terms.

Decision rules are expressions in the form *"if... then..."*, in symbols $\Phi \to \Psi$. Examples of decision rules are shown below:

(*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, high*) $\to$ (*Temperature, high*)

(*Solar Energy, high*) $\to$ (*Temperature, high*)

Formally the language of decision rules, called a *decision language*, is defined as follows. Let $S = (U, A)$ be an information system. With every $B \subseteq A$ we associate a formal language, i.e., a set of formulas $For(B)$. Formulas of $For(B)$ are built up from attribute-value pairs (a, v) where $a \in B$ and $v \in V_a$ by means of logical connectives $\wedge$ (*and*), $\vee$ (*or*), $\sim$ (*not*), in the standard way.

For any $\Phi \in For(B)$ by $||\Phi||_S$ we denote the set of all objects $x \in U$ satisfying Φ in S and refer to $||\Phi||_S$ as the *meaning* of Φ in S.

The meaning $||\Phi||_S$ of Φ in S is defined inductively as follows:

$$\begin{aligned} ||(a,v)||_S &= \{x \in U : a(x) = v\} \text{ for all } a \in B \text{ and } v \in V_a,\ ||\Phi \vee \Psi||_S \\ &= ||\Phi||_S \cup ||\Psi||_S, ||\Phi \wedge \Psi||_S = ||\Phi||_S \cap ||\Psi||_S, || \sim \Phi||_S \\ &= U - ||\Phi||_S\ . \end{aligned} \tag{7.4}$$

A formula Φ is *true* in S if $||\Phi||_S = U$.

A *decision rule* in S is an expression $\Phi \to \Psi$, read *if* Φ *then* Ψ, where $\Phi \in For(C)$, $\Psi \in For(D)$ and C, D are condition and decision attributes, respectively; Φ and Ψ are referred to as *conditions* and *decisions* of the rule, respectively.

A decision rule $\Phi \to \Psi$ is *true* in S if $||\Phi||_S \subseteq ||\Psi||_S$.

The number $supp_S(\Phi, \Psi) = card(||\Phi \wedge \Psi||_S)$ will be called the *support* of the rule $\Phi \to \Psi$ in S. We consider a probability distribution $p_U(x) = 1/card(U)$ for $x \in U$ where U is the (non-empty) universe of objects of S; we have $p_U(X) = card(X)/card(U)$ for $X \subseteq U$. For any formula Φ we associate its probability in S defined by

$$\pi_S(\Phi) = p_U(||\Phi||_S). \tag{7.5}$$

With every decision rule $\Phi \to \Psi$ we associate a conditional probability

$$\pi_S(\Psi|\Phi) = p_U(\|\Psi\|_S|\ \|\Phi\|_S) \tag{7.6}$$

that Ψ is true in S given Φ is true in S, which we call the *certainty factor*. We have

$$\pi_S(\Psi|\Phi) = \frac{card(||\Phi \wedge \Psi||_S)}{card(||\Phi||_S)}\ , \tag{7.7}$$

where $||\Phi||_S \neq \emptyset$.

This coefficient is now widely used in data mining and is called the *confidence coefficient*.

Obviously, $\pi_S(\Psi|\Phi) = 1$ if and only if $\Phi \to \Psi$ is true in S.

If $\pi_S(\Psi|\Phi) = 1$, then $\Phi \to \Psi$ will be called a *certain decision* rule; if $0 < \pi_S(\Psi|\Phi) < 1$ the decision rule will be referred to as a *uncertain decision* rule.

Besides, we will also use a *coverage factor* defined by

$$\pi_S(\Phi|\Psi) = p_U(\|\Phi\|_S|\ \|\Psi\|_S)\ , \tag{7.8}$$

which is the conditional probability that Φ is true in S, given Ψ is true in S with the probability $\pi_S(\Psi)$. Obviously, we have

$$\pi_S(\Phi|\Psi) = \frac{card(||\Phi \wedge \Psi||_S)}{card(||\Psi||_S)}. \tag{7.9}$$

The certainty factors in S can be also interpreted as the frequency of objects having the property Ψ in the set of objects having the property Φ and the coverage factor – as the frequency of objects having the property Φ in the set of objects having the property Ψ.

The number

$$\sigma_S(\Phi, \Psi) = \frac{supp_S(\Phi, \Psi)}{card(U)} = \pi_S(\Psi|\Phi) \cdot \pi_S(\Phi) \tag{7.10}$$

will be called the *strength* of the decision rule $\Phi \to \Psi$ in S.

For example, for the decision rule

(*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, low*) $\to$ (*Temperature, high*),

support = 90, *strength* = 0.25, *certainty* = 0.74, *coverage* = 0.43.

Summing up, decision rules, which are in fact logical implications, constitute a logical counterpart of approximations: certain rules correspond to the lower approximation, whereas the uncertain rules correspond to the boundary region. Thus we have two formal tools to deal with vagueness: approximations and implications. Mathematically, approximations are basic operations (interior and closure) in a topology generated by a data set. Thus if we want to prove properties of the data (find patterns in the data) the topological language of approximations is the right tool. However, in order to describe the patterns in the data for practical use, the logical language of implications is the proper one.

The certainty and the coverage factors of decision rules are conditional probabilities which express how exact our knowledge (data) is about the considered reality. Let us satisfy ourselves that the factors are not assumed arbitrarily, but are computed from the data, and are thus in a certain sense objective.

From the logical point of view, the certainty factor can be interpreted as a degree of truth of the decision rule, i.e., how strongly the decision can be trusted in view of the data. On the contrary, the coverage factor can be viewed as a degree of truth of the "inverted" decision rule, i.e., to what degree the reasons for a decision can be trusted in view of the data.

Statistically, the certainty factor simply reveals the frequency of facts satisfying conditions, among the facts satisfying decision of the decision rule, whereas the interpretation of the coverage factor is converse.

Finally, let us briefly comment upon the concept of the strength of a decision rule. This number simply expresses the ratio of all facts that can be classified by the decision rule to all facts in the data table. It will be shown in the next sections that this coefficient plays an essential role in further considerations, and it will be used to reformulate Bayes' theorem.

7.5 Decision Algorithms

In this section we define the notion of a decision algorithm, which is a logical counterpart of a decision table. Informally, a decision algorithm is a set of mutually exclusive and exhaustive decision rules associated with a given decision table. An example of a decision algorithm associated with *Table 7.1* is given below.

1. (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, high*) $\rightarrow$ (*Temperature, high*).
2. (*Solar Energy, high*) $\rightarrow$ (*Temperature, high*).
3. (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, low*) $\rightarrow$ (*Temperature, high*).

4. (*Solar Energy, low*) $\rightarrow$ (*Temperature, low*).

5. (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, low*) $\rightarrow$ (*Temperature, low*).

Formally, a decision algorithm is defined as follows.

Let $Dec(S) = \{\Phi_i \rightarrow \Psi_i\}_{i=1}^{m}$, $m \geq 2$, be a set of decision rules associated with a decision table $S = (U, C, D)$.

a. If for every $\Phi \rightarrow \Psi, \Phi' \rightarrow \Psi' \in Dec(S)$ we have $\Phi = \Phi'$ or $||\Phi \wedge \Phi'||_S = \emptyset$, and $\Psi = \Psi'$ or $||\Psi \wedge \Psi'||_S = \emptyset$, then we will say that $Dec(S)$ is the set of pairwise *mutually exclusive* (*independent*) decision rules in S.

b. If $||\bigvee_{i=1}^{m} \Phi_i||_S = U$ and $||\bigvee_{i=1}^{m} \Psi_i||_S = U$ we will say that the set of decision rules $Dec(S)$ *covers U*.

c. If $\Phi \rightarrow \Psi \in Dec(S)$ and $supp_S(\Phi, \Psi) \neq 0$ we will say that the decision rule $\Phi \rightarrow \Psi$ is *admissible* in S.

d. If $\bigcup_{X \in U/D} C_*(X) = ||\bigvee_{\Phi \rightarrow \Psi \in Dec^+(S)} \Phi||_S$ where $Dec^+(S)$ is the set of all certain decision rules from $Dec(S)$, we will say that the set of decision rules $Dec(S)$ preserves the *consistency* of the decision table $S = (U, C, D)$.

The set of decision rules $Dec(S)$ that satisfies a), b), c), and d) (i.e., is independent); covers U; preserves the consistency of S and all decision rules $\Phi \rightarrow \Psi \in Dec(S)$ are admissible in S – will be called a *decision algorithm* in S.

If $\Phi \rightarrow \Psi$ is a decision rule then the decision rule $\Psi \rightarrow \Phi$ will be called an *inverse* decision rule of $\Phi \rightarrow \Psi$.

Let $Dec^*(S)$ denote the set of all inverse decision rules of $Dec(S)$.

It can be shown that $Dec^*(S)$ satisfies a), b), c), and d), i.e., it is a decision algorithm in S.

If $Dec(S)$ is a decision algorithm then $Dec^*(S)$ will be called an *inverse* decision algorithm of $Dec(S)$.

The inverse decision algorithm for the decision algorithm 1) - 5) is as follows:

1') (*Temperature, high*) $\rightarrow$ (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, high*).

2') (*Temperature, high*) $\rightarrow$ (*Solar Energy, high*).

3') (*Temperature, high*) $\rightarrow$ (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, low*).

4') (*Temperature, low*) $\rightarrow$ (*Solar Energy, low*).

5') (*Temperature, low*) $\rightarrow$ (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, low*).

The inverse decision algorithm can be used as an *explanation* of a decision in terms of conditions, i.e., it gives *reasons* for decisions.

As mentioned at the beginning of this section, a decision algorithm is a counterpart of a decision table. The properties a), b), c), and d) have been chosen in such a way that the decision algorithm preserves basic properties of the data in the decision table, in particular approximations and boundary regions of decisions.

A crucial issue in rough-set-based data analysis is the generation of "optimal" decision algorithms from the data. This is a complex task, particularly when large data bases are concerned. Many methods and algorithms have been proposed to deal with this problem, but I will not dwell upon this issue here, for I intend to restrict this chapter to rudiments of rough set theory only. The interested reader is advised to consult the references and the Web (see end of this chapter for a web-site listing).

7.6 Decision Algorithms and Approximations

Decision algorithms can be used as a formal language for describing approximations.

For example, *certain* decision rules

1. (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, high*) $\rightarrow$ (*Temperature, high*) and

2. (*Solar Energy, high*) $\rightarrow$ (*Temperature, high*)

describe the lower approximation of the decision (*Temperature, high*). The uncertain decision rule

3. (*Solar Energy, medium*) $\wedge$ (*Volcanic Activity, low*) $\rightarrow$ (*Temperature, high*)

describes the boundary region of the decision (*Temperature, high*).

The above relationships can be defined more precisely as follows.

Let $Dec(S)$ be a decision algorithm in S and let $\Phi \rightarrow \Psi \in Dec(S)$. By $C(\Psi)$ we denote the set of all conditions of Ψ in $Dec(S)$ and by $D(\Phi)$, the set of all decisions of Φ in $Dec(S)$.

Then we have the following relationships:

$$C_*(||\Psi||_S) = ||\bigvee_{\Phi' \in C(\Psi),\ \pi(\Psi|\Phi')=1} \Phi'||_S \,, \tag{7.11}$$

$$C^*(||\Psi||_S) = ||\bigvee_{\Phi' \in C(\Psi),\ 0<\pi(\Psi|\Phi')\leq 1} \Phi'||_S \,, \tag{7.12}$$

$$BN_C(||\Psi||_S) = ||\bigvee_{\Phi' \in C(\Psi),\ 0<\pi(\Psi|\Phi')<1} \Phi'||_S\ . \tag{7.13}$$

From the above properties we can get the following definitions:

i. From Equation (7.11), if $||\Phi||_S = C_*(||\Psi||_S)$, then formula Φ will be called the *C-lower approximation* of the formula Ψ and will be denoted by $C_*(\Psi)$.

ii. From Equation (7.12), if $||\Phi||_S = C^*(||\Psi||_S)$, then the formula Φ will be called the *C-upper approximation* of the formula Φ and will be denoted by $C^*(\Psi)$.

iii. From Equation (7.13), if $||\Phi||_S = BN_C(||\Psi||_S)$, then Φ will be called the *C-boundary* of the formula Ψ and will be denoted by $BN_C(\Psi)$.

The above properties say that any decision $\Psi \in Dec(S)$ can be uniquely described by the following certain and uncertain decision rules, respectively:

$$C_*(\Psi) \rightarrow \Psi,$$

$$BN_C(\Psi) \rightarrow \Psi.$$

Thus, decision algorithms can be viewed as a logical counterpart of approximations, or more exactly as a formal language to describe approximations. The language of decision rules is more convenient to describe decisions in terms of conditions than the topological language of approximations. However, approximations give better insight into vagueness and uncertainty of data.

7.7 Some Properties of Decision Algorithms

Decision algorithms have interesting probabilistic properties, which are discussed next.

Let $Dec(S)$ be a decision algorithm and let $\Phi \rightarrow \Psi \in Dec(S)$. Then the following properties are valid:

$$\sum_{\Phi' \in C(\Psi)} \pi_S(\Phi'|\Psi) = 1 \tag{7.14}$$

$$\sum_{\Psi' \in D(\Phi)} \pi_S(\Psi'|\Phi) = 1 \tag{7.15}$$

Table 7.2. Certainty and coverage factors for the decision algorithm 1) – 5).

Decision rule	Support	Strength	Certainty	Coverage
1	20	0.06	1.00	0.10
2	100	0.27	1.00	0.47
3	90	0.25	0.74	0.43
4	120	0.33	1.00	0.79
5	34	0.09	0.26	0.21

$$\pi_S(\Psi) = \sum_{\Phi' \in C(\Psi)} \pi_S(\Psi|\Phi') \cdot \pi_S(\Phi') = \sum_{\Phi' \in C(\Psi)} \sigma_S(\Phi', \Psi) \tag{7.16}$$

$$\pi_S(\Phi|\Psi) = \frac{\pi_S(\Psi|\Phi) \cdot \pi_S(\Phi)}{\sum_{\Phi' \in C(\Psi)} \pi_S(\Psi|\Phi') \cdot \pi_S(\Phi')} = \frac{\sigma_S(\Phi, \Psi)}{\pi_S(\Psi)} \tag{7.17}$$

That is, any decision algorithm, and consequently any decision table, satisfies (7.14), (7.15), (7.16), and (7.17). Observe that (7.16) is the well-known *Total Probability Theorem* and (7.17) is the *Bayes' Theorem*.

Note that we are not referring to prior and posterior probabilities – fundamental in Bayesian data analysis philosophy. The Bayes' Theorem in our case says that: if an implication $\Phi \rightarrow \Psi$ is true in the degree $\pi_S(\Psi|\Phi)$, then the inverse implication $\Psi \rightarrow \Phi$ is true in the degree $\pi_S(\Phi|\Psi)$.

In other words the Bayes' Theorem in our case reveals some relationships between decisions and their reasons, or – more exactly – it discovers some relationships in every set of data.

Thus in order to compute the certainty and coverage factors of decision rules it is enough to know the strength (support) of all decision rules in the decision algorithm only.

The certainty and coverage factors for the decision algorithm 1) – 5) are given in *Table 7.2*.

The strength of decision rules can be computed from the data or can be a subjective assessment.

From the certainty factors of the decision algorithm we can conclude the following:

1. If the solar energy is medium and the volcanic activity is high then the temperature is *certainly* high.

2. If the solar energy is high then the temperature is *certainly* high.

3. If the solar energy is medium and the volcanic activity is low then the *probability* that the temperature is high equals 0.74.

4. If the solar energy is low then the temperature is *certainly* low.

5. If the solar energy is medium and the volcanic activity is low then the *probability* that the temperature is low equals 0.26.

The coverage factors of the decision algorithm lead us to the following explanation of global warming:

1') If the temperature is high then the *probability* that the solar energy is medium and the volcanic activity is high amounts to 0.10.

2') If the temperature is high then the *probability* that the solar energy is high equals 0.47.

3') If the temperature is high then the *probability* that the solar energy is medium and the volcanic activity is low equals 0.43.

4') If the temperature is low then the *probability* that the solar energy is low equals 0.79.

5') If the temperature is low then the *probability* that the solar energy is medium and the volcanic activity is low equals 0.21.

Summing up, from the data we can conclude that:

- Medium solar energy and high volcanic activity or high solar energy *certainly* cause high temperature.
- Low solar energy *certainly* causes low temperature.
- Medium solar energy and low volcanic activity cause:
  - high temperature with (probability = 0.74) and
  - low temperature with (probability = 0.26).

Whereas the data lead to the following explanation of global warming.
The reasons for high temperature are:

- Medium solar energy and high volcanic activity (probability = 0.10).
- High solar energy (probability = 0.47).
- Medium solar energy and low volcanic activity (probability = 0.43).

The reasons for low temperature are:

- Low solar energy (probability = 0.79).
- Medium solar energy and low volcanic activity (probability = 0.21).

In short, we can derive from the data the following conclusions:

- Medium solar energy and high volcanic activity or high solar energy *certainly* cause high temperature.
- Low solar energy *certainly* causes low temperature.
- Medium solar energy and low volcanic activity *most probably* cause high temperature.

and the following explanations:

- The *most probable* reason for high temperature is high solar energy.
- The *most probable* reason for low temperature is low solar energy.

Summing up, from the rough-set view, Bayes' theorem reveals probabilistic structure of a data set (i.e., any decision table or decision algorithm) without referring to either prior or posterior probabilities, inherently associated with Bayesian statistical inference methodology. In other words, it identifies probabilistic relationships between conditions and decisions in decision algorithms, in contrast to classical Bayesian reasoning, where data are employed to verify prior probabilities. This is not the case in rough-set-based data analysis.

Let us also stress that Bayes' theorem in the rough-set approach has a new mathematical form based on strength of decision rules, which essentially simplifies computations and gives us a new look at the theorem.

7.8 Conclusions

Approximations, basic concepts of rough-set theory, have been defined and discussed. Some probabilistic properties of approximation have been revealed, in particular the relationship with the Total Probability Theorem and the Bayes' Theorem. These relationships give a new efficient method to draw conclusions from data, without referring to prior and posterior probabilities intrinsically associated with Bayesian reasoning. The application of the proposed method has been outlined, by means of a simple tutorial example concerning global warming.

Acknowledgment

Thanks are due to Professor Hirotaka Nakayama for critical remarks.

References

Düntsch, I., and Günter, G., 2000, *Rough Set Data Analysis: A Road to Non-Invasive Knowledge Discovery*, Metoδos Publishers, Banglor, UK, and Bissendorf, Germany.

Flinkman, M., Michalowski, W., Nilsson, S., Slowinski R., Susmaga R., and Wilk, S., to appear, Use of rough sets analysis to classify Siberian forest ecosystems according to net primary production of phytomass, *INFOR 2000*, **35**(3):145–160.

Grzymała–Busse, D.M., and Grzymała–Busse, J.W., 1994, Comparison of machine learning and knowledge acquisition methods of rule induction based on rough sets, in W. Ziarko, ed., *Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93): Workshops in Computing*, Springer-Verlag and British Computer Society, London, UK and Berlin, Germany, pp. 282–288.

Grzymała–Busse, J.W., 1994, ESEP: An expert system for environmental protection, in W. Ziarko, ed., *Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93): Workshops in Computing*, Springer-Verlag and British Computer Society, London, UK and Berlin, Germany, pp. 466–474.

Grzymała–Busse, J.W., and Gunn, J.D., 1995, Global temperature analysis based on the rule induction system LERS, in M. Dąbrowski and M. Michalewicz, eds, *Proceedings of the Fourth International Workshop on Intelligent Information Systems*, Augustów, Poland, June 5–9, Institute of Computer Science, Polish Academy of Sciences, Warsaw, pp. 148–158.

Gunn, J.D., and Grzymała–Busse, J.W., 1994, Global temperature stability by rule induction: An interdisciplinary bridge, *Human Ecology*, **22**:59–81.

Keiser, K., Szladow, A., and Ziarko, W., 1992, Rough sets theory applied to a large multispecies toxicity database, in *Proceedings of the Fifth International Workshop on QSAR in Environmental Toxicology*, Natural Resources Research Institute, University of Minnesota, Duluth, MN, USA.

Pal, S.K., and Skowron, A., eds, 1999, *Rough Fuzzy Hybridization*, Springer-Verlag, Berlin, Germany and New York, NY, USA.

Pawlak, Z., 1982, Rough sets, *International Journal of Computer and Information Sciences*, **11**:341–356.

Pawlak, Z., 1991, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Boston, MA, USA; London, UK; and Dordrecht, Netherlands.

Polkowski, L., and Skowron, A., eds, 1998, *Rough Sets and Current Trends in Computing: Lecture Notes in Artificial Intelligence 1424*, Springer-Verlag, Berlin, Germany and New York, NY, USA.

Polkowski, L., and Skowron, A., eds, 1998, *Rough Sets in Knowledge Discovery*, Vol. 1-2, Physica Verlag, A Springer Company, Heidelberg, Germany.

Polkowski, L., Tsumoto, S., and Lin, T.Y., eds, 2000, *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, Physica-Verlag, Heidelberg, Germany and New York, NY, USA, a Springer-Verlag Company.

Reinhard, A., Stawski, B., Weber, T., and Wybraniec–Skardowska, U., 1992, An application of rough set theory in the control conditions on a polder, in R. Słowiński, ed., *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 331–362.

Teghem, J., and Charlet, J.-M., 1992, Use of 'rough sets' method to draw premonitory factors for earthquakes by emphasizing gas geochemistry: The case of a low seismic activity context in Belgium, in R. Słowiński, ed., *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 165–180.

Zhong, N., Skowron, A., and Ohsuga, S., eds, 1999, *New Direction in Rough Sets, Data Mining, and Granular-Soft Computing*, Springer-Verlag, Berlin, Germany and New York, NY, USA.

More Information About Rough Sets Can be Found at:

http://www.cs.uregina.ca/roughset
http://www.infj.ulst.ac.uk/staff/I.Duentsch
http://www-idss.cs.put.poznan.pl/staff/slowinski/
http://alfa.mimuw.edu.pl/logic/
http://www.idi.ntnu.no/aleks/rosetta/
http://www.infj.ulst.ac.uk/cccz23/grobian/grobian.html

Chapter 8

Generalized DEA Model for Multiple Criteria Decision Making

Ye Boon Yun, Hirotaka Nakayama, Masao Arakawa, and Hiroshi Ishikawa

Abstract

DEA(Data Envelopment Analysis) is now widely applied for evaluating relative efficiencies of decision making units (DMUs) performing similar tasks in a production system that consumes multiple inputs to produce multiple outputs. So far, several DEA models have been developed: the CCR model, the BCC model, and the FDH model are well known as basic DEA models. These models can be considered from two viewpoints: one based on the domination structure in the primal form, and the other characterized by a determination of the production possibility set in the dual form.

On the other hand, MCDA (Multiple Criteria Decision Analysis) has been studied as a way to help decision makers (DMs) come to their final decisions in MCDM (Multiple Criteria Decision Making) problems. One of the main tasks in this research is how to incorporate the value judgments of DMs into decision support systems. If DMs can make their decisions by seeing the efficiencies (or inefficiencies) of alternatives, the idea of DEA can be applied to MCDM problems. In this event, it is important to know what value judgment the domination structure of each

DEA model reflects. Moreover, a model that can treat a wide range of DMs' value judgments is required. To this end, we propose here a generalized DEA model and discuss its practical use in MCDM problems.

Keywords: Data envelopment analysis, multiple criteria decision making, CCR model, BCC model, FDH model, generalized DEA model.

8.1 Introduction

DEA (Data Envelopment Analysis) was suggested by Charnes, Cooper, and Rhodes (CCR), and built on Farrell's (1957) idea, which is concerned with the estimation of technical efficiency and efficient frontiers. The CCR model (Charnes *et al*., 1978, 1979) generalized the single output/single input ratio efficiency measure for each DMU (Decision Making Unit) to multiple outputs/multiple inputs situations by forming the ratio of a weighted sum of outputs to a weighted sum of inputs. DEA is a method for measuring the relative efficiency of DMUs performing similar tasks in a production system that consumes multiple inputs to produce multiple outputs. The main characteristics of DEA are that:

1. It can be applied to analyze multiple outputs and multiple inputs without pre-assigned weights.
2. It can be used for measuring a relative efficiency based on the observed data without knowing information on the production function.
3. It can incorporate decision makers' (DMs') preferences.

Later, Banker, Charnes, and Cooper (BCC) suggested a model for distinguishing between technical efficiency and scale inefficiency in DEA. The BCC model (Banker *et al*., 1984) relaxed the constant-returns-to-scale assumption of the CCR model and made it possible to investigate whether the performance of each DMU was conducted in a region of increasing, constant, or decreasing returns to scale in situations of multiple outputs and multiple inputs. In addition, Tulkens (1993) introduced a relative efficiency to the non-convex free disposable hull (FDH) of the observed data defined by Deprins *et al*. (1984), and formulated a mixed-integer programming to calculate the relative efficiency for each DMU. In addition to the basic models mentioned above, several extended models have been studied. Examples include the cone ratio model (Charnes *et al.*, 1989), the polyhedral cone ratio model (Charnes *et al*., 1990), Seiford and Thrall's model (1990), Wei and Yu's model (1997), and so on.

Relationships between DEA and multiple criteria decision analysis (MCDA) have been studied from several viewpoints by many authors. Belton (1992), and

Belton and Vickers (1993) measured efficiency as a weighted sum of input and output. Stewart (1996) showed the equivalence between the CCR model and a linear value function model for multiple outputs and multiple inputs. Joro *et al.* (1998) proved structural correspondences between DEA models and multiple objective linear programming (MOLP) using an achievement scalarizing function proposed by Wierzbicki (1980). In particular, various ways of introducing preference information into DEA formulations have been developed. Golany (1988) suggested a so-called target-setting model, which allows DMs to select the preferred set of output levels given the input levels of a DMU. Thanassoulis and Dyson (1992) introduced models that can be used to estimate alternative output and input levels, in order to render relatively inefficient DMUs efficient. Zhu (1996) proposed a model that calculates efficiency scores incorporating the DMs' preference information, whereas Korhonen (1997) applied an interactive technique to progressively reveal preferences. Halme *et al.* (1999) evaluated an efficiency of DMU in terms of pseudo-concave value function, by considering a tangent cone of the feasible set at the DM's most preferred solution. Agrell and Tind (1998) showed correspondences among the CCR model (Charnes *et al*., 1978), the BCC model (Banker *et al.*, 1984), and the FDH model (Tulkens, 1993) and MCDA model according to the property of a partial Lagrangean relaxation. Yun *et al*. (2000) suggested a concept of "value-free efficiency" in the observed data.

In this study, we propose a generalized model for DEA – the so-called GDEA model – which can treat basic DEA models, specifically, the CCR model, the BCC model, and the FDH model in a unified way. In addition, we show theoretical properties of relationships among the GDEA model and the previously mentioned DEA models. The GDEA model makes it possible to calculate the efficiency of DMUs incorporating various preference structures of DMs. Finally, we suggest a dual approach GDEA$_D$ to GDEA and show also that GDEA$_D$ can reveal domination relations among all DMUs. The rest of this chapter is organized as follows. Section 8.2 presents the notations used here, as well as brief explanations on basic DEA models. Section 8.3 discusses multiple criteria decision making (MCDM). Section 8.4 proposes the GDEA model based on a parametric domination. Section 8.5 presents a dual approach to GDEA, that is, the GDEA$_D$ model based on a production possibility set. In Section 8.6, we compare the efficiency of GDEA and several DEA models for each DMU through illustrative examples. Conclusions are presented in Section 8.7.

8.2 Basic DEA Models

In the following discussion, we assume that there exist n DMUs to be evaluated. Each DMU consumes varying amounts of m different inputs to produce p

different outputs. Specifically, DMUj consumes amounts $\boldsymbol{x}_j := (x_{ij})$ of inputs $(i = 1, \cdots, m)$ and produces amounts $\boldsymbol{y}_j := (y_{kj})$ of outputs $(k = 1, \cdots, p)$. For these constants, which generally take the form of observed data, we assume $x_{ij} > 0$ for each $i = 1, \cdots, m$ and $y_{kj} > 0$ for each $k = 1, \cdots, p$. Further, we assume that there are no duplicated units in the observed data. The $p \times n$ output matrix for the n DMUs is denoted by $\boldsymbol{Y}$, and the $m \times n$ input matrix for the n DMUs is denoted by $\boldsymbol{X}$. $\boldsymbol{x}_o := (x_{1o}, \cdots, x_{mo})$ and $\boldsymbol{y}_o := (y_{1o}, \cdots, y_{po})$ are amounts of inputs and outputs of DMUo, which is evaluated. In addition, ε is a small positive number ("non-Archimedean") and $\mathbf{1} = (1, \cdots, 1)$ is a unit vector.[1]

For convenience, the following notations for vectors in $\mathbb{R}^{p+m}$ will be used:

$$\boldsymbol{z}_o > \boldsymbol{z}_j \iff z_{io} > z_{ij}, \;\; i = 1, \cdots, p+m,$$

$$\boldsymbol{z}_o \geqq \boldsymbol{z}_j \iff z_{io} \geqq z_{ij}, \;\; i = 1, \cdots, p+m,$$

$$\boldsymbol{z}_o \geq \boldsymbol{z}_j \iff z_{io} \geqq z_{ij}, \;\; i = 1, \cdots, p+m \text{ but } \boldsymbol{z}_o \neq \boldsymbol{z}_j.$$

So far, several DEA models have been developed. Among them, the CCR model (Charnes *et al.*, 1978; 1979), the BCC model (Banker *et al.*, 1984), and the FDH model (Tulkens, 1993) are well known as basic DEA models. These models are based on the domination structure in the primal form, and moreover these are characterized by how to determine the production possibility set in the dual form; the convex cone, the convex hull, and the free disposable hull for the observed data, respectively. These models are further discussed in the following subsections.

8.2.1 The CCR model

The CCR model, which was suggested by Charnes *et al.* (1978), is a fractional linear programming problem. It can be solved by being transformed into an equivalent linear programming problem. Therefore, the primal problem (CCR) with an input-oriented model can be formulated as the following:

$$\begin{array}{lll} \underset{\mu_k,\, \nu_i}{\text{maximize}} & \displaystyle\sum_{k=1}^{p} \mu_k y_{ko} & \text{(CCR)} \\ \text{subject to} & \displaystyle\sum_{i=1}^{m} \nu_i x_{io} = 1, & \\ & \displaystyle\sum_{k=1}^{p} \mu_k y_{kj} - \sum_{i=1}^{m} \nu_i x_{ij} \leqq 0, \;\; j = 1, \cdots, n, & \\ & \mu_k \geqq \varepsilon, \;\; \nu_i \geqq \varepsilon, \;\; k = 1, \cdots, p\,; \; i = 1, \cdots, m. & \end{array}$$

[1] Archimedean property : If $x \in \mathbb{R}$, $y \in \mathbb{R}$ and $x > 0$, then there exists a positive integer n such that $nx > y$. Non-Archimedean ε is a small positive number not satisfying Archimedean property.

The dual problem (CCR_D) to the problem (CCR) is given by

$$\begin{aligned} \underset{\theta,\, \boldsymbol{\lambda},\, \boldsymbol{s}_x,\, \boldsymbol{s}_y}{\text{minimize}} \quad & \theta - \varepsilon(\mathbf{1}^T \boldsymbol{s}_x + \mathbf{1}^T \boldsymbol{s}_y) \\ \text{subject to} \quad & \boldsymbol{X}\boldsymbol{\lambda} - \theta \boldsymbol{x}_o + \boldsymbol{s}_x = \mathbf{0}, \\ & \boldsymbol{Y}\boldsymbol{\lambda} - \boldsymbol{y}_o - \boldsymbol{s}_y = \mathbf{0}, \\ & \boldsymbol{\lambda} \geqq \mathbf{0},\ \ \boldsymbol{s}_x \geqq \mathbf{0},\ \ \boldsymbol{s}_y \geqq \mathbf{0}, \\ & \theta \in \mathbb{R},\ \ \boldsymbol{\lambda} \in \mathbb{R}^n,\ \ \boldsymbol{s}_x \in \mathbb{R}^m, \boldsymbol{s}_y \in \mathbb{R}^p. \end{aligned} \qquad (\mathrm{CCR}_D)$$

The ‘efficiency’ in the CCR model is introduced as follows:

Definition 1. (CCR-efficiency) A DMUo is *CCR-efficient* if and only if the optimal value $\sum_{k=1}^{p} \mu_k^* y_{ko}$ to the problem (CCR) equals one. Otherwise, the DMUo is said to be *CCR-inefficient*.

Definition 2. (CCR_D-efficiency) A DMUo is *CCR_D-efficient* if and only if for the optimal solution $(\theta^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{s}_x^*,\ \boldsymbol{s}_y^*)$ to the problem (CCR_D), the following two conditions are satisfied:

(i) θ^* is equal to one;

(ii) the slack variables $\boldsymbol{s}_x^*$ and $\boldsymbol{s}_y^*$ are all zero.

Otherwise, the DMUo is *CCR_D-inefficient*.

Note that the above two definitions are equivalent due to the well known duality of linear programming.

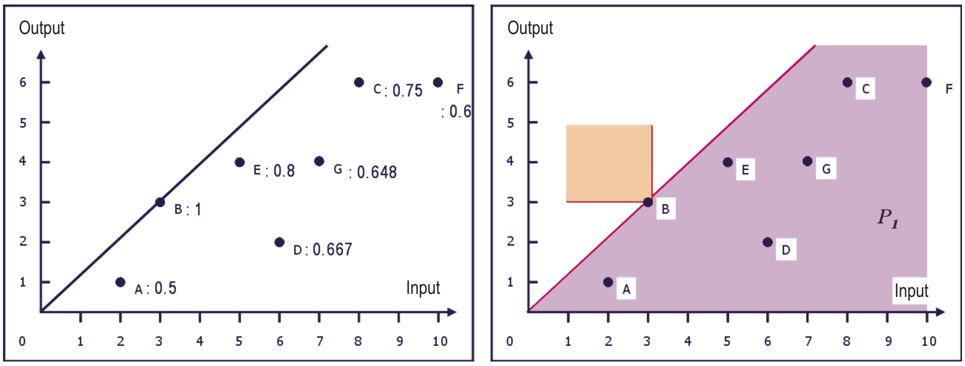
Additionally, the production possibility set P_1 in the dual form of the CCR model is the *convex cone* (or conical hull) generated by the observed data, which implies that the scale efficiency of a DMU is constant, that is to say, it involves constant returns to scale. Therefore, P_1 can be denoted by

$$P_1 = \Big\{ (\boldsymbol{y}, \boldsymbol{x}) \mid Y\boldsymbol{\lambda} \geqq \boldsymbol{y},\ X\boldsymbol{\lambda} \leqq \boldsymbol{x},\ \boldsymbol{\lambda} \geqq \mathbf{0} \Big\}.$$

and the definition of CCR-efficiency (or CCR_D-efficiency) can be transformed into the following:

Definition 3. DMUo is said to be *Pareto efficient in* P_1 if and only if there does not exist $(\boldsymbol{y}, \boldsymbol{x}) \in P_1$ such that $(\boldsymbol{y}, -\boldsymbol{x}) \geq (\boldsymbol{y}_o, -\boldsymbol{x}_o)$.

It is readily seen that the Pareto efficiency in P_1 is equivalent to the CCR-efficiency. *Figure 8.1* shows a geometric interpretation on the relation between the primal form of CCR model and the dual one.



(a) Primal form (b) Dual form

Figure 8.1. CCR efficient frontier and production possibility set generated by the CCR model from the observed data.

8.2.2 The BCC model

The BCC model of Banker *et al.* (1984) is formulated similarly to that for the CCR model. The dual problem for the BCC model is obtained by adding the convexity constraint $\mathbf{1}^T\boldsymbol{\lambda} = 1$ to the dual problem (CCR$_D$) and thus, the variable u_o appears in the primal problem. The efficiency degree of a DMUo with respect to the BCC model can be measured by solving the problem.

$$\begin{aligned}
\underset{\mu_k,\,\nu_i,\,u_o}{\text{maximize}} \quad & \sum_{k=1}^{p} \mu_k y_{ko} - u_o && \text{(BCC)}\\
\text{subject to} \quad & \sum_{i=1}^{m} \nu_i x_{io} = 1, \\
& \sum_{k=1}^{p} \mu_k y_{kj} - \sum_{i=1}^{m} \nu_i x_{ij} - u_o \leqq 0, \quad j = 1, \cdots, n, \\
& \mu_k \geqq \varepsilon, \quad \nu_i \geqq \varepsilon, \quad k = 1, \cdots, p\,; \; i = 1, \cdots, m.
\end{aligned}$$

The dual problem (BCC$_D$) to the problem (BCC) is formulated as follows:

$$\begin{aligned}
\underset{\theta,\,\boldsymbol{\lambda},\,\boldsymbol{s}_x,\,\boldsymbol{s}_y}{\text{minimize}} \quad & \theta - \varepsilon(\mathbf{1}^T\boldsymbol{s}_x + \mathbf{1}^T\boldsymbol{s}_y) && \text{(BCC}_D\text{)}\\
\text{subject to} \quad & \boldsymbol{X\lambda} - \theta\boldsymbol{x}_o + \boldsymbol{s}_x = \mathbf{0}, \\
& \boldsymbol{Y\lambda} - \boldsymbol{y}_o - \boldsymbol{s}_y = \mathbf{0}, \\
& \mathbf{1}^T\boldsymbol{\lambda} = 1, \\
& \boldsymbol{\lambda} \geqq \mathbf{0}, \quad \boldsymbol{s}_x \geqq \mathbf{0}, \quad \boldsymbol{s}_y \geqq \mathbf{0}, \\
& \theta \in \mathbb{R}, \quad \boldsymbol{\lambda} \in \mathbb{R}^n, \quad \boldsymbol{s}_x \in \mathbb{R}^m, \boldsymbol{s}_y \in \mathbb{R}^p.
\end{aligned}$$

The ‘efficiency’ in the BCC model is given by the following two definitions which are equivalent to each other due to the duality of linear programming.

Definition 4. (BCC-efficiency) A DMUo is *BCC-efficient* if and only if the optimal value $\left(\sum_{k=1}^{p} \mu_k^* y_{ko} - u_o^*\right)$ to the problem (BCC) equals one. Otherwise, the DMUo is said to be *BCC-inefficient*.

Definition 5. (BCC$_D$-efficiency) A DMUo is *BCC$_D$-efficient* if and only if for an optimal solution $(\theta^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{s}_x^*,\ \boldsymbol{s}_y^*)$ to the problem (BCC$_D$), the following two conditions are satisfied:

(i) θ^* is equal to one;

(ii) the slack variables $\boldsymbol{s}_x^*$ and $\boldsymbol{s}_y^*$ are all zero.

Otherwise, the DMUo is said to be *BCC$_D$-inefficient*.

The presence of the constraint $\mathbf{1}^T\boldsymbol{\lambda} = 1$ in the dual problem (BCC$_D$) yields that the production possibility set P_2 in the BCC model is the *convex hull* generated by the observed data. Therefore, P_2 can be obtained as

$$P_2 = \left\{ (\boldsymbol{y}, \boldsymbol{x}) \mid Y\boldsymbol{\lambda} \geqq \boldsymbol{y},\ X\boldsymbol{\lambda} \leqq \boldsymbol{x},\ \mathbf{1}^T\boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda} \geqq \mathbf{0} \right\}.$$

and the definition of BCC$_D$-efficiency can be transformed into the following:

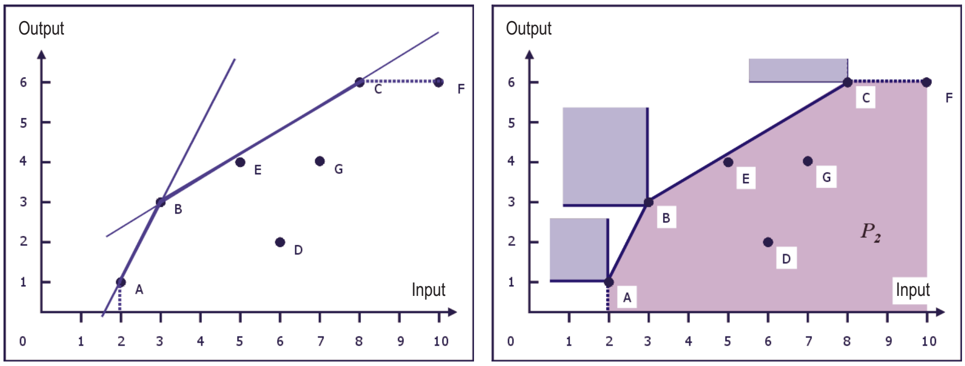
Definition 6. DMUo is said to be *Pareto efficient in* P_2 if and only if there does not exist $(\boldsymbol{y}, \boldsymbol{x}) \in P_2$ such that $(\boldsymbol{y}, -\boldsymbol{x}) \geq (\boldsymbol{y}_o, -\boldsymbol{x}_o)$.

It is readily seen that the Pareto efficiency in P_2 is equivalent to the BCC-efficiency. *Figure 8.2* shows a geometric interpretation on the relation between the primal form of BCC model and the dual one.

8.2.3 The FDH model

The FDH model by Tulkens (1993) is formulated as follows:

$$\begin{aligned}
\underset{\theta,\,\boldsymbol{\lambda},\,\boldsymbol{s}_x,\,\boldsymbol{s}_y}{\text{minimize}} \quad & \theta - \varepsilon(\mathbf{1}^T\boldsymbol{s}_x + \mathbf{1}^T\boldsymbol{s}_y) \qquad (\text{FDH}_D)\\
\text{subject to} \quad & \boldsymbol{X\lambda} - \theta\boldsymbol{x}_o + \boldsymbol{s}_x = \mathbf{0},\\
& \boldsymbol{Y\lambda} - \boldsymbol{y}_o - \boldsymbol{s}_y = \mathbf{0},\\
& \mathbf{1}^T\boldsymbol{\lambda} = 1;\ \ \lambda_j \in \{0,\ 1\}\ \text{ for each }\ j = 1, \cdots, n,\\
& \boldsymbol{\lambda} \geqq \mathbf{0},\ \ \boldsymbol{s}_x \geqq \mathbf{0},\ \ \boldsymbol{s}_y \geqq \mathbf{0},\\
& \theta \in \mathbb{R},\ \ \boldsymbol{\lambda} \in \mathbb{R}^n,\ \ \boldsymbol{s}_x \in \mathbb{R}^m, \boldsymbol{s}_y \in \mathbb{R}^p.
\end{aligned}$$



(a) Primal form (b) Dual form

Figure 8.2. BCC efficient frontier and production possibility set generated by the BCC model from the observed data.

Here, however, it is seen that the problem (FDH$_D$) is a mixed integer programming problem, and hence the traditional linear optimization methods cannot apply to it. An optimal solution is obtained by means of a simple vector comparison procedure at the end.

For a DMUo, the optimal solution θ^* to the problem (FDH$_D$) is equal to the value R_o^* defined by

$$R_o^* = \min_{j \in D(o)} \max_{i=1,\cdots m} \left\{ \frac{x_{ij}}{x_{io}} \right\}, \tag{8.1}$$

where $D(o) = \{\, j \mid \boldsymbol{x}_j \leq \boldsymbol{x}_o \text{ and } \boldsymbol{y}_j \geqq \boldsymbol{y}_o, \ \ j = 1, \cdots, n \,\}$.

R_o^* is substituted for θ^* as the efficiency degree for DMUo in the FDH model. The 'efficiency' in the FDH model is given by the following.

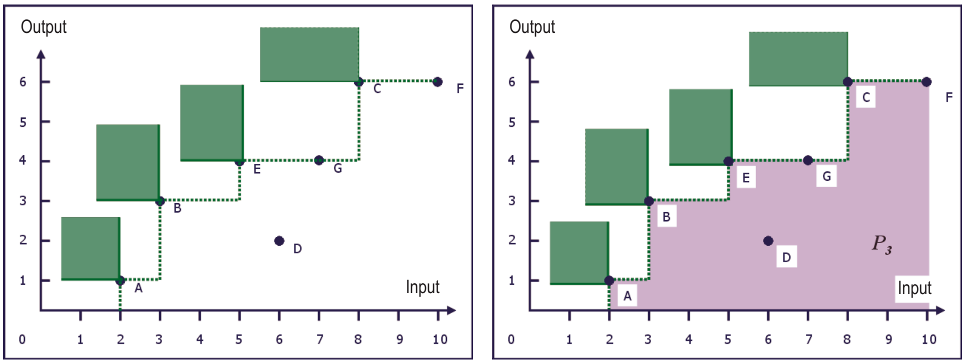
Definition 7. (FDH-efficiency) A DMUo is *FDH-efficient* if and only if R_o^* equals to one. If $R_o^* < 1$, the DMUo is said to be *FDH-inefficient*.

Definition 8. (FDH$_D$-efficiency) A DMUo is *FDH$_D$-efficient* if and only if for an optimal solution $(\theta^*, \ \boldsymbol{\lambda}^*, \ \boldsymbol{s}_x^*, \ \boldsymbol{s}_y^*)$ to the problem (FDH$_D$), the following two conditions are satisfied:

(i) θ^* is equal to one;

(ii) the slack variables $\boldsymbol{s}_x^*$ and $\boldsymbol{s}_y^*$ are all zero.

Otherwise, the DMUo is said to be *FDH$_D$-inefficient*.



(a) Primal form (b) Dual form

Figure 8.3. FDH efficient frontier and production possibility set generated by the FDH model from the observed data.

It can be seen that the above two definitions are equivalent to each other, and the production possibility set P_3, which is a free disposable hull, is given by

$$P_3 = \left\{ (\boldsymbol{y}, \boldsymbol{x}) \mid Y\boldsymbol{\lambda} \geqq \boldsymbol{y},\ X\boldsymbol{\lambda} \leqq \boldsymbol{x},\ \mathbf{1}^T\boldsymbol{\lambda} = 1,\ \lambda_j \in \{0,1\},\ j = 1, \cdots, n \right\}. \tag{8.2}$$

Besides, the definition of FDH-efficiency (or FDH$_D$-efficiency) can be transformed into the following:

Definition 9. DMUo is said to be *Pareto efficient in* P_3 if and only if there does not exist $(\boldsymbol{y}, \boldsymbol{x}) \in P_3$ such that $(\boldsymbol{y}, -\boldsymbol{x}) \geq (\boldsymbol{y}_o, -\boldsymbol{x}_o)$.

It is shown that the Pareto efficiency in P_3 is equivalent to the FDH-effciency. *Figure 8.3* shows a geometric interpretation on the relation between the primal form of FDH model and the dual one.

8.3 Multiple Criteria Decision Making

Consider decision making problems with multiple criteria $f_1, \cdots, f_r$ which are to be maximized. Let S denote the set of alternatives. For this problem, $\boldsymbol{x}^o \in S$ or $\boldsymbol{f}^o (= \boldsymbol{f}(\boldsymbol{x}^o))$ is said to be *Pareto efficient* if and only if there does not exist $\boldsymbol{x} \in S$ such that $\boldsymbol{f}(\boldsymbol{x}) \geq \boldsymbol{f}(\boldsymbol{x}^o)$. Usually, a Pareto efficient solution is not necessarily uniquely determined, but there are several Pareto efficient solutions. In practical decision making, therefore, we have to determine a solution among the Pareto efficient solutions. To this end, the value judgments of DMs are introduced. The multi-attribute utility (value) analysis provides some mathematical form for these value judgments of DMs. On the other hand, interactive multi-objective programming techniques search a decision making solution eliciting partial information on the DMs' value judgments. In any case, the final solution strongly depends on the value judgment.

The idea of DEA can be applied to MCDM problems, if a final decision making solution is not necessarily determined by the method itself but done by seeing efficiencies (or inefficiencies) of alternatives. Let DMUs be identified with alternatives in MCDM problems. Then, it should be noted that efficiencies in DEA also depend on value judgments. It should be emphasized that the ratio of output to input is merely one of these value judgments. In many production activity analyses, the ratio of output to input is naturally adopted as one such value judgment. In applying DEA to a wide range of practical problems, however, there are some cases in which the ratio value judgement is not adequate. In other words, in some cases a DMU is not necessarily judged to be inefficient, even though a CCR model shows it to be so.

The additive value may be represented by a linear weighted sum of each criterion. In this circumstance, a value judgment is reflected by a set of weights to criteria. If a DMU maximizes a weighted sum of criteria, it can be regarded as efficient in terms of the value judgment. Therefore, a DMU can be said to be additive value efficient if it maximizes a weighted sum of criteria. The set of additive value efficient DMUs is identical to the set of efficient DMUs under the BCC model [or the additive model of DEA by Charnes *et al.* (1985)].

Depending on the situation, value judgments of DMs cannot necessarily be represented by a weighted sum of criteria. Nonlinear value functions can be used for more general value judgments of DMs [e.g., pseudo-concave value functions by Halme *et al.* (1999)]. The notion of efficiency without introducing any value judgment is the Pareto efficiency. We call this "the value free efficiency." The set of value free efficient DMUs is identical to that of the FDH model. In the following sections, we describe a generalized DEA model which embeds these value judgments in a unified model. The key idea of the model is to introduce a domination structure with one parameter varying from the value free structure to a ratio value structure.

8.4 GDEA Based on Parametric Domination Structure

In this section, we formulate a GDEA model based on a domination structure and define a new 'efficiency' in the GDEA model. Next, we establish relationships between the GDEA model and the basic DEA models mentioned in Section 8.2.

We formulate a generalized DEA model by employing the augmented Tcheby-shev scalarizing function (Sawaragi *et al.*, 1985). The GDEA model, which can evaluate efficiencies in several basic models as special cases, is the following:

$$\begin{aligned}
\underset{\Delta,\,\mu_k,\,\nu_i}{\text{maximize}} \quad & \Delta \qquad\qquad \text{(GDEA)}\\
\text{subject to} \quad & \Delta \leqq \tilde{d}_j + \alpha\left(\sum_{k=1}^{p}\mu_k(y_{ko}-y_{kj}) + \sum_{i=1}^{m}\nu_i(-x_{io}+x_{ij})\right),\\
& \quad j = 1,\cdots,n,\\
& \sum_{k=1}^{p}\mu_k + \sum_{i=1}^{m}\nu_i = 1,\\
& \mu_k,\ \nu_i \geqq \varepsilon,\ \ k=1,\cdots,p\,;\, i=1,\cdots,m,
\end{aligned}$$

where $\tilde{d}_j = \max\limits_{\substack{k=1,\cdots,p\\ i=1,\cdots,m}} \{\mu_k(y_{ko}-y_{kj}),\ \nu_i(-x_{io}+x_{ij})\}$ and α is a positive number.

Note that when $j = o$, the right-hand side of the inequality constraint in the problem (GDEA) is zero, and hence its optimal value is not greater than zero. We define 'efficiency' in the GDEA model as follows.

Definition 10. (α-efficiency) For a given positive number α, DMUo is defined to be α*-efficient* if and only if the optimal value to the problem (GDEA) is equal to zero. Otherwise, DMUo is said to be α*-inefficient*.

8.4.1 Relationships between GDEA and DEA

In this subsection, we establish theoretical properties for relationships among efficiencies in the basic DEA models and those in the GDEA model.

Theorem 1. *DMUo is FDH-efficient if and only if DMUo is α-efficient for some sufficiently small positive number α.*

Theorem 2. *DMUo is BCC-efficient if and only if DMUo is α-efficient for some sufficiently large positive number α.*

Table 8.1. An example of 1-input and 1-output.

DMU	A	B	C	D	E	F
Input	2	3	4.5	4	6	5.5
Output	1	3	3.5	2	5	4.0

Consider the problem (GDEA′) in which the constraint $\sum_{k=1}^{p} \mu_k y_{ko} = \sum_{i=1}^{m} \nu_i x_{io}$ is added to the problem (GDEA):

$$
\text{(GDEA}'\text{)} \quad \underset{\Delta,\,\mu_k,\,\nu_i}{\text{maximize}} \quad \Delta
$$
$$
\text{subject to} \quad \Delta \leqq \tilde{d}_j + \alpha\left(\sum_{k=1}^{p} \mu_k(y_{ko} - y_{kj}) + \sum_{i=1}^{m} \nu_i(-x_{io} + x_{ij})\right), \quad j = 1, \cdots, n,
$$
$$
\sum_{k=1}^{p} \mu_k y_{ko} - \sum_{i=1}^{m} \nu_i x_{io} = 0,
$$
$$
\sum_{k=1}^{p} \mu_k + \sum_{i=1}^{m} \nu_i = 1,
$$
$$
\mu_k, \;\; \nu_i \geqq \varepsilon, \;\; k = 1, \cdots, p\,; \, i = 1, \cdots, m,
$$

where $\tilde{d}_j = \max\limits_{\substack{k=1,\cdots,p \\ i=1,\cdots,m}} \{\mu_k(y_{ko} - y_{kj}), \; \nu_i(-x_{io} + x_{ij})\}$ and α is a given positive number.

Theorem 3. *DMUo is CCR-efficient if and only if DMUo is α-efficient for some sufficiently large positive α when regarding the problem (GDEA) as the problem (GDEA′).*

From the stated theorems, it is seen that the CCR-efficiency, BCC-efficiency, and FDH-efficiency for each DMU can be evaluated by varying the parameter α in the problem (GDEA).

8.4.2 An illustrative example

Here, we explain the α-efficiency in the GDEA model with a simple illustrative example and reveal domination relations among all DMUs by GDEA.

Assume that there are six DMUs which consume one input to produce one output, as seen in *Table 8.1*.

Table 8.2 shows the results of efficiency in the basic DEA models and α-efficiency in the GDEA model. In the upper half of *Table 8.2*, we see that a DMU is efficient if the optimal value is equal to one in the CCR model, the BCC model, and the FDH models, respectively. The lower half of *Table 8.2* shows the α-efficiency by changing a parameter α. It can be seen that if $\alpha = 0.1$, the α-efficiency of

Table 8.2. The optimal values in basic DEA models and GDEA model.

DMU	A	B	C	D	E	F
CCR model	0.50	**1.00**	0.78	0.50	0.83	0.73
BCC model	**1.00**	**1.00**	0.83	0.63	**1.00**	0.75
FDH model	**1.00**	**1.00**	**1.00**	0.75	**1.00**	**1.00**
(i) $\alpha = 10$ (GDEA′)	–9.33	**0.00**	–3.25	–11.33	–0.73	–3.74
(ii) $\alpha = 10$	**0.00**	**0.00**	–2.10	–11.00	**0.00**	–3.35
(iii) $\alpha = 3$	**0.00**	**0.00**	**0.00**	–4.00	**0.00**	–0.55
(iv) $\alpha = 1$	**0.00**	**0.00**	**0.00**	–2.00	**0.00**	**0.00**
(v) $\alpha = 0.1$	**0.00**	**0.00**	**0.00**	–1.10	**0.00**	**0.00**

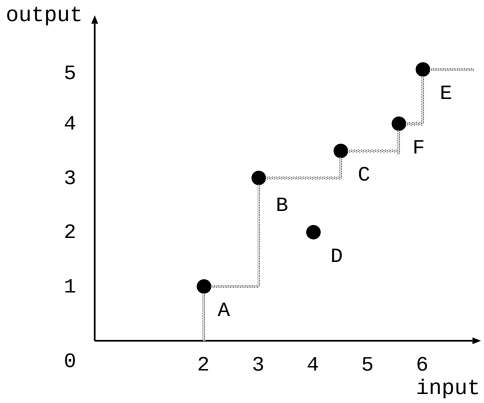


Figure 8.4. Efficient frontier generated by GDEA model with $\alpha \fallingdotseq 0$.

each DMU is the same as the FDH-efficiency. If $\alpha = 10$, the α-efficiency of each DMU is the same as the BCC-efficiency, and, moreover, if $\alpha = 10$ in the problem (GDEA′), then the α-efficiency is equivalent to the CCR-efficiency. Furthermore, *Figure 8.4*, *Figure 8.5*, and *Figure 8.6* represent the efficient frontier generated by varying α in the GDEA model.

This example shows that by varying the value of parameter α, various efficiencies of the basic DEA models can be measured in a unified way on the basis of this GDEA model. Furthermore, the relationships among efficiencies for these models become transparent.

8.5 GDEA Based on Production Possibility

In this section, we consider a dual approach to the GDEA model introduced in Section 8.3. We formulate a GDEA$_D$ model based on the production possibility

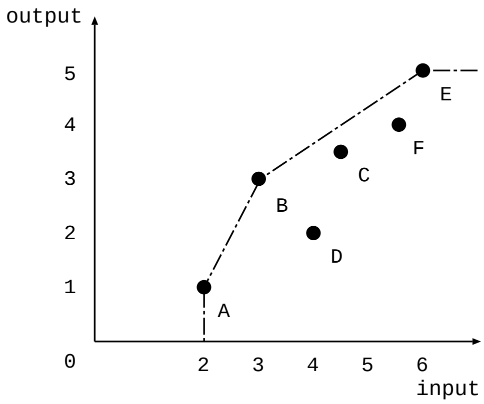


Figure 8.5. Efficient frontier generated by GDEA model with $\alpha = 10$.

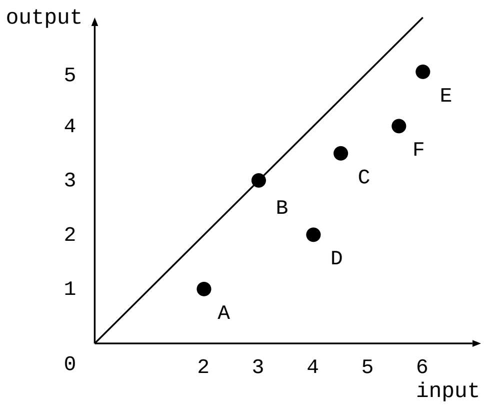


Figure 8.6. Efficient frontier generated by GDEA$'$ model with $\alpha = 10$.

set and define 'efficiency' in the GDEA$_D$ model. Next, we establish relationships between the GDEA$_D$ model and dual models of the basic DEA models mentioned in Section 8.2.

To begin with, an output-input vector $\boldsymbol{z}_j$ of a DMUj, $j = 1, \cdots, n$ and output-input matrix Z of all DMUs respectively, denoted by

$$\boldsymbol{z}_j := \begin{pmatrix} \boldsymbol{y}_j \\ -\boldsymbol{x}_j \end{pmatrix},\ j = 1, \cdots, n \ \text{ and } \ Z := \begin{pmatrix} \boldsymbol{Y} \\ -\boldsymbol{X} \end{pmatrix}.$$

In addition, we denote a $(p+m) \times n$ matrix Z_o by $Z_o := (\boldsymbol{z}_o, \cdots, \boldsymbol{z}_o)$, where o is the index of DMU to be evaluated.

The production possibility sets in the CCR model, the BCC model, and the FDH model in Section 8.2 are reformulated as follows:

$$P_1' = \{\boldsymbol{z} \mid Z\boldsymbol{\lambda} \geqq \boldsymbol{z},\ \boldsymbol{\lambda} \geqq \boldsymbol{0}\}$$
$$P_2' = \left\{\boldsymbol{z} \mid Z\boldsymbol{\lambda} \geqq \boldsymbol{z},\ \boldsymbol{1}^T\boldsymbol{\lambda} = 1,\ \boldsymbol{\lambda} \geqq \boldsymbol{0}\right\}$$
$$P_3' = \left\{\boldsymbol{z} \mid Z\boldsymbol{\lambda} \geqq \boldsymbol{z},\ \boldsymbol{1}^T\boldsymbol{\lambda} = 1,\ \lambda_j \in \{0,1\},\ j = 1,\cdots,n\right\}$$

and the 'efficiencies' in these models are redefined.

Definition 11. DMUo is said to be *Pareto efficient in* P_1' if and only if there does not exist $(\boldsymbol{y}, -\boldsymbol{x}) \in P_1'$ such that $(\boldsymbol{y}, -\boldsymbol{x}) \geq (\boldsymbol{y}_o, -\boldsymbol{x}_o)$.

Definition 12. DMUo is said to be *Pareto efficient in* P_2' if and only if there does not exist $(\boldsymbol{y}, -\boldsymbol{x}) \in P_2'$ such that $(\boldsymbol{y}, -\boldsymbol{x}) \geq (\boldsymbol{y}_o, -\boldsymbol{x}_o)$.

Definition 13. DMUo is said to be *Pareto efficient in* P_3' if and only if there does not exist $(\boldsymbol{y}, -\boldsymbol{x}) \in P_3'$ such that $(\boldsymbol{y}, -\boldsymbol{x}) \geq (\boldsymbol{y}_o, -\boldsymbol{x}_o)$.

Remark 1. (Joro *et al*., 1998) *Here, Definitions 11-13 correspond to the CCR-efficiency (or CCR$_D$-efficiency), BCC-efficiency (or BCC$_D$-efficiency), and the FDH-efficiency (or FDH$_D$-efficiency), respectively.*

The dual problem to (GDEA′) introduced in Section 8.4 is formulated as follows:

$$\begin{aligned} \underset{\omega,\,\kappa,\,\boldsymbol{\lambda},\,\boldsymbol{s_z}}{\text{minimize}} \quad & \omega - \varepsilon \boldsymbol{1}^T \boldsymbol{s_z} \qquad\qquad (\text{GDEA}_D) \\ \text{subject to} \quad & \{\alpha(Z_o - Z) + D_{\boldsymbol{z}}\}\boldsymbol{\lambda} - \boldsymbol{\omega} + \boldsymbol{s_z} + \kappa \boldsymbol{z}_o = \boldsymbol{0} \\ & \boldsymbol{1}^T\boldsymbol{\lambda} = 1, \\ & \boldsymbol{\lambda} \geqq \boldsymbol{0},\ \ \boldsymbol{s_z} \geqq \boldsymbol{0}, \end{aligned}$$

where $\boldsymbol{\omega} = (\omega, \cdots, \omega)$ and α is a given positive number. A $(p+m) \times n$ matrix $D_{\boldsymbol{z}} := (\boldsymbol{d}_1, \cdots, \boldsymbol{d}_n)$ is a matrix $(Z - Z_o)$. It is replaced by 0, except for the maximal component (if there exist plural maximal components, only one is chosen from among them) in each row. Especially, it is seen that when κ is fixed at 0, (GDEA$_D$) becomes the dual problem to (GDEA), since κ is the dual variable to the second constraint in (GDEA′).

We define an 'efficiency' for a DMUo in the GDEA$_D$ model:

Definition 14. (α_D-efficiency) For a given positive α, DMUo is said to be α_D-*efficient* if and only if the optimal solution $(\omega^*,\ \kappa^*,\ \boldsymbol{\lambda}^*,\ \boldsymbol{s}_z^*)$ to the problem (GDEA$_D$) satisfies the following two conditions:

(i) ω^* is equal to zero;

(ii) the slack variable $\boldsymbol{s}_z^*$ is zero.

Otherwise, DMUo is said to be α_D-*inefficient*.

It should be noted particularly that for an optimal solution $(\omega^*, \kappa^*, \boldsymbol{\lambda}^*, \boldsymbol{s}_{\boldsymbol{z}}^*)$ to the problem GDEA$_D$, ω^* is not greater than zero because of the strong duality of (GDEA) and (GDEA$_D$) in a linear programming problem, and the 'non-Archimedean' property of ε.

8.5.1 Relationships between GDEA$_D$ and DEA

In this subsection, we summarize theoretical properties of relationships among efficiencies in basic DEA models and the GDEA$_D$ model.

Theorem 4. *Let κ be fixed at 0 in (GDEA$_D$). DMUo is Pareto efficient in P_3' if and only if DMUo is α_D-efficient for some sufficiently small positive number α.*

Theorem 5. *Let κ be fixed at 0 in (GDEA$_D$). DMUo is Pareto efficient in P_2' if and only if DMUo is α_D-efficient for some sufficiently large positive number α.*

Theorem 6. *DMUo is Pareto efficient in P_1' if and only if DMUo is α_D-efficient for some sufficiently large positive number α.*

8.5.2 Optimal solutions to (GDEA$_D$)

In this subsection, we explain the meaning of optimal solutions ω^*, $\boldsymbol{\lambda}^*$, $\boldsymbol{s}_{\boldsymbol{z}}^*$ to (GDEA$_D$). ω^* gives a measure of relative efficiency for DMUo. In other words, it represents the degree to which DMUo is inefficient; that is, how far DMUo is from the efficient frontier generated with the given α. $\boldsymbol{\lambda}^* := (\lambda_1^*, \cdots, \lambda_n^*)$ represents a domination relation between DMUo and another DMU. That is, it means that the DMUo is dominated by DMUj if λ_j for some $j \neq 0$ is positive. $\boldsymbol{s}_x^*$ represents the slack of inputs and $\boldsymbol{s}_y^*$ is the surplus of outputs for performance of the DMUo.

Consider an illustrative example as shown in *Table 8.3*. The table shows the results of the CCR-efficiency, BCC-efficiency, and FDH-efficiency, respectively, in the example. *Table 8.4* shows the optimal solution $(\omega^*, \kappa^*, \boldsymbol{\lambda}^*, \boldsymbol{s}_{\boldsymbol{z}}^*)$ to (GDEA$_D$) ($\varepsilon = 10^{-6}$) when α is given as 10^{-6} and κ is fixed at 0. *Table 8.5* shows the optimal solution $(\omega^*, \kappa^*, \boldsymbol{\lambda}^*, \boldsymbol{s}_{\boldsymbol{z}}^*)$ to (GDEA$_D$) ($\varepsilon = 10^{-6}$) when α is given by 10 and κ is fixed at 0. Finally, *Table 8.6* shows the optimal solution $(\omega^*, \kappa^*, \boldsymbol{\lambda}^*, \boldsymbol{s}_{\boldsymbol{z}}^*)$ to (GDEA$_D$) ($\varepsilon = 10^{-6}$) when α is given as 10.

Here, we can see that the FDH-efficiency, BCC-efficiency, and CCR-efficiency are equivalent to the α-efficiency with $\alpha = 10^{-6}$ $(\kappa = 0)$, $\alpha = 10$ $(\kappa = 0)$ and $\alpha = 10$ $(\text{nonfixed } \kappa)$, respectively, from the result of *Table 8.4*, *Table 8.5*, and

Table 8.3. An Example of 1-input and 1-output and optimal value in the problems (CCR), (BCC) and (FDH).

DMU	Input	Output	CCR model	BCC model	FDH model
A	2	1	0.5	1	1
B	3	3	1	1	1
C	8	6	0.75	1	1
D	6	2	0.333	0.417	0.5
E	5	4	0.8	0.933	1
F	10	6	0.6	$1 - 2 \times 10^{-6}$	0.8
G	7	4	0.571	0.667	0.714

Table 8.4. Optimal solution to (GDEA$_D$) with $\alpha = 10^{-6}$ and fixed $\kappa = 0$.

DMU	ω^*	$\boldsymbol{\lambda}^*$	$\boldsymbol{s}_{\boldsymbol{z}}^* = (\boldsymbol{s}_x^*,\ \boldsymbol{s}_y^*)$
A	0	$\lambda_A^* = 1$	(0,0)
B	0	$\lambda_B^* = 1$	(0,0)
C	0	$\lambda_C^* = 1$	(0,0)
D	-0.5	$\lambda_B^* = \lambda_E^* = 0.5$	(0,0)
E	0	$\lambda_E^* = 1$	(0,0)
F	0	$\lambda_C^* = 1$	(2,0)
G	0	$\lambda_E^* = 1$	(2,0)

Table 8.5. Optimal solution to (GDEA$_D$) with $\alpha = 10$ and fixed $\kappa = 0$.

DMU	ω^*	$\boldsymbol{\lambda}^*$	$\boldsymbol{s}_{\boldsymbol{z}}^* = (\boldsymbol{s}_x^*,\ \boldsymbol{s}_y^*)$
A	0	$\lambda_A^* = 1$	(0,0)
B	0	$\lambda_B^* = 1$	(0,0)
C	0	$\lambda_C^* = 1$	(0,0)
D	-7.803	$\lambda_B^* = 0.765, \lambda_C^* = 0.235$	(0,0)
E	-0.441	$\lambda_B^* = 0.631, \lambda_C^* = 0.369$	(0,0)
F	0	$\lambda_C^* = 1$	(20,0)
G	-8.281	$\lambda_B^* = 0.378, \lambda_C^* = 0.622$	(0,0)

Table 8.6. Optimal solution to (GDEA$_D$) with $\alpha = 10$ and non-fixed κ.

DMU	ω^*	$\boldsymbol{\lambda}^*$	$\boldsymbol{s}_{\boldsymbol{z}}^* = (\boldsymbol{s}_x^*,\ \boldsymbol{s}_y^*)$	κ^*
A	-11.333	$\lambda_C^* = 1$	(0,0)	38.667
B	0	$\lambda_B^* = 1$	(0,0)	0
C	-2.571	$\lambda_B^* = 1$	(0,0)	-5.929
D	-24.500	$\lambda_C^* = 1$	(0,0)	7.750
E	-2.778	$\lambda_B^* = 1$	(0,0)	-3.444
F	-7.500	$\lambda_C^* = 1$	(0,0)	-1.250
G	-8.727	$\lambda_C^* = 1$	(0,0)	2.818

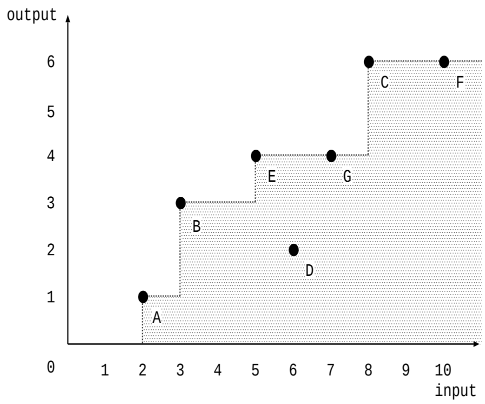


Figure 8.7. Efficient frontier generated by GDEA$_D$ model with $\alpha = 10^{-6}$ and fixed $\kappa = 0$.

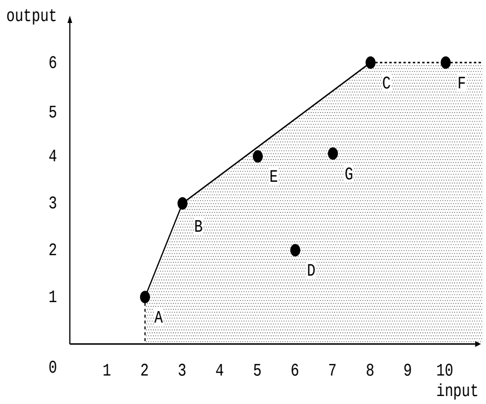


Figure 8.8. Efficient frontier generated by GDEA$_D$ model with $\alpha = 10$ and fixed $\kappa = 0$.

Table 8.6 and *Figure 8.7*, *Figure 8.8*, and *Figure 8.9*. In other words, the FDH-efficiency, BCC-efficiency, and CCR-efficiency can be obtained by changing the parameter α in the GDEA$_D$ model.

Now, we interpret a meaning of optimal solutions $(\omega^*, \kappa^*, \boldsymbol{\lambda}^*, \boldsymbol{s}_{\boldsymbol{z}}^*)$ to (GDEA$_D$). Note that ω^* gives a measure of relative efficiency for DMUo. In other words, it represents the degree to which DMUo is inefficient, that is, how far DMUo is from the efficient frontier generated with the given α.

$\boldsymbol{\lambda}^* := (\lambda_1^*, \cdots, \lambda_n^*)$ represents a domination relation between DMUo and another DMU. That is, it means that the DMUo is dominated by DMUj if λ_j for some $j \neq o$ is positive. For example, as is seen in *Table 8.4*, the optimal solution for the DMU D is $\lambda_B^* = 0.5$ and $\lambda_E^* = 0.5$, and hence DMU D is dominated by DMU B

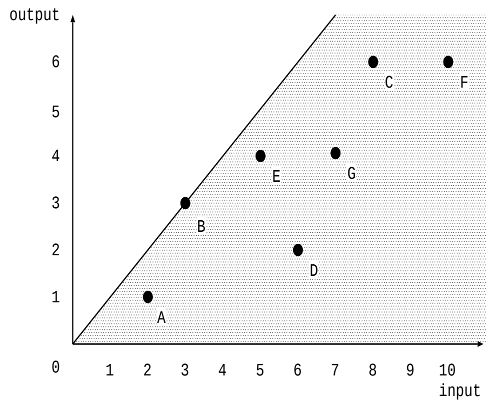


Figure 8.9. Efficient frontier generated by GDEA$_D$ model with $\alpha = 10$ and non-fixed κ.

and DMU E (see *Figure 8.7*). In addition, in *Table 8.5*, the optimal solution for the DMU E is $\lambda_B^* = 0.631$ and $\lambda_C^* = 0.369$, and hence DMU E is dominated by linear combination of DMU B and DMU C (see *Figure 8.8*). As is seen in *Table 8.6*, the optimal solution for the DMU C is $\lambda_B^* = 1$, and hence DMU D is dominated by a point on the line through DMU B and the origin (see *Figure 8.9*).

$\boldsymbol{s}_x^*$ represents the slack of inputs and $\boldsymbol{s}_y^*$ is the surplus of outputs for performance of the DMUo. For instance, DMU G has the optimal solution $\omega^* = 0,\ \lambda_E^* = 1$ and $(\boldsymbol{s}_{\boldsymbol{x}}^*,\ \boldsymbol{s}_{\boldsymbol{y}}^*) = (2,\ 0)$. DMU G is α-inefficient because $\boldsymbol{s}_x^*$ is not equal to zero although $\omega^* = 0$. It implies that DMU G has the larger surplus amount of input than DMU E with the same output.

8.6 Comparison Between GDEA and DEA Models

Now, we compare the efficiency in basic DEA models and GDEA model for the data in Taylor *et al.* (1997). We have data from 13 Mexican commercial banks over two years (1990–1991) from Taylor's group. As shown in *Table 8.7*, each bank has the total income as the single output. Total income is the sum of a bank's interest and non-interest income. Total deposits and total non-interest expense are the two inputs used to generate the output. Interest income includes interest earned from loan activities. Total non-interest income includes dividends, fees, and other non-interest revenue. The total deposits input variable includes the bank's interest paying deposit liabilities. Total non-interest expense includes personnel and administrative costs, commissions paid, banking support fund contributions and other non-interest operating costs. Thus, we evaluate the efficiency for each bank with the annual data, that is, consider α-efficiency corresponding to several values

Table 8.7. Input and output values for 13 Mexican banks, 1990–1991 (billions of nominal pesos).

	1990			1991		
Bank	Deposits	Non-int. expense	Int. income plus non-int. income	Deposits	Non-int. expense	Int. income plus non-int. income
(1) Banamex	35,313.90	2,500.88	14,247.10	57,510.90	3,670.33	15,764.60
(2) Bancomer	34,504.60	2,994.70	12,682.10	59,965.00	3,872.40	15,877.00
(3) Serfin	30,558.20	1,746.50	11,766.40	46,987.20	2,709.20	12,694.10
(4) Intermac	7,603.53	1,011.40	3,422.40	13,458.00	1,165.20	4,212.20
(5) Cremi	1,977.18	1,628.80	2,889.10	5,108.97	760.60	2,102.70
(6) Bancreser	2,405.00	140.70	1,050.50	3,314.32	190.80	1,681.10
(7) MercNort	2,146.06	338.30	1,320.10	3,714.72	463.30	1,377.40
(8) BCH	2,944.00	260.8	1,410.00	3,728.00	402.90	1,794.10
(9) Confia	1,962.34	266.60	1,568.00	3,324.43	364.90	1,944.40
(10) Bancen	1,815.73	196.70	946.20	2,544.96	242.70	848.80
(11) Promex	1,908.23	251.30	1,162.80	3,080.00	320.40	1,251.40
(12) Banoro	1,372.78	169.60	598.20	2,799.00	224.40	810.50
(13) Banorie	488.17	71.90	340.80	680.88	86.80	373.00

Source: Taylor *et al.* (1997).

$\alpha = 0.1,\ 0.5,\ 1,\ 10,\ 15$ (only 1991) and 10^3. *Table 8.8* represents the results of analyses under the basic DEA models and the GDEA model.

As shown in the tables, the GDEA model with $\alpha = 0.1$ provides FDH efficiency. It means that there is no change in α-efficient DMUs for smaller α than 0.1. In addition, the GDEA model with $\alpha = 10$ yields BCC efficiency in *Table 8.8*, while $\alpha = 15$ does in *Table 8.9*. Also, there is no change in α-efficiency of DMUs, even if taking greater α than 10 or 15.

Moreover, CCR efficiency can be figured by taking α sufficiently large in the GDEA model and adding the constraint $\boldsymbol{x}_o^T \nu = \boldsymbol{y}_o^T \mu$. This operation shows that the number of efficient DMUs decreases as a parameter α increases in general. Particularly, note the α-efficiency for $\alpha = 0.5$ and $\alpha = 1$. This represents an intermediate efficiency between FDH-efficiency and BCC-efficiency. In practice, there are decision-making problems which cannot correspond to a special value judgment such as "ratio value efficiency"[2] in the CCR model, "sum value efficiency"[3] in the BCC model, and so on. In contrast to the existing DEA models, the GDEA model can incorporate various value judgments of DMs by changing a parameter α, and then several kinds of efficiency of the basic DEA models can be measured in a unified way on the basis of the GDEA model. Furthermore, the relationships among efficiencies for these models become transparent by considering GDEA.

8.7 Conclusions

In this paper, we suggested the GDEA model based on parametric domination structure, and defined α-efficiency in the GDEA model. In addition, we investigated theoretical properties of relationships between the GDEA model and existing DEA models, specifically, the CCR model, the BCC model, and the FDH model. It was then proved that the GDEA model makes it possible to evaluate efficiencies of several DEA models in a unified way, and to incorporate various preference structures of DMs. Through a numerical example, it has been shown that the mutual relations among all decision-making units can be grasped by varying α in GDEA model. Furthermore, we proposed the GDEA$_D$ model based on production possibility as a dual approach to GDEA, and defined α_D-efficiency in the GDEA$_D$ model. Also, we clarified the relations between the GDEA$_D$ model and existing DEA dual models, and interpreted the meaning of an optimal value to the problem (GDEA$_D$). As a result, it is possible to make a quantitative analysis for inefficiency on the basis of surplus of inputs and slack of outputs. Moreover, through an illustrative example, it

[2] We named the CCR-efficiency *ratio value efficiency*, because the ratio of the weighted sum of outputs to the weighted sum of inputs is maximized by the CCR model (see Yun *et al.*, 2000).

[3] We named the BCC-efficiency *sum value efficiency*, because the difference of the weighted sum of outputs and the weighted sum of inputs is maximized by the BCC model (see Yun *et al.*, 2000).

Table 8.8. DEA Mexican bank analysis, 13 banks, 1990. Output is total interest and non-interest income; inputs are total deposits and non-interest expense.

Bank	CCR θ	CCR Class	BCC θ	BCC RTS	GDEA $\alpha = 10^3$ $(\boldsymbol{x}_o^T \nu = \boldsymbol{y}_o^T \mu)$	GDEA $\alpha = 10$	GDEA $\alpha = 1$	GDEA $\alpha = 0.5$	GDEA $\alpha = 0.1$
(1) Banamex	0.816	NE	1.000	D	–123.46	0.00	0.00	0.00	0.00
(2) Bancomer	0.646	NE	0.890	–	–744.67	–7,282.88	–358.41	0.00	0.00
(3) Serfin	0.902	NE	1.000	D	–11.88	0.00	0.00	0.00	0.00
(4) Intermac	0.573	NE	0.809	–	–285.50	–1,648.99	0.00	0.00	0.00
(5) Cremi	1.000	E	1.000	C	0.00	0.00	0.00	0.00	0.00
(6) Bancreser	1.000	E	1.000	C	0.00	0.00	0.00	0.00	0.00
(7) MercNort	0.750	NE	0.757	–	–126.73	–1,078.91	–149.92	–102.55	–19.69
(8) BCH	0.829	NE	0.837	–	–70.89	–390.60	–11.27	–0.08	0.00
(9) Confia	1.000	E	1.000	C	0.00	0.00	0.00	0.00	0.00
(10) Bancen	0.778	NE	0.803	–	–94.29	–390.09	–8.06	0.00	0.00
(11) Promex	0.782	NE	0.797	–	–79.50	–506.79	–29.08	–6.76	0.00
(12) Banoro	0.588	NE	0.644	–	–299.20	–606.52	–12.81	0.00	0.00
(13) Banorie	0.862	NE	1.000	I	–58.55	0.00	0.00	0.00	0.00

C = constant returns to scale; D = decreasing returns to scale (RTS); E = efficient; I = increasing returns to scale; NE = not efficient.

Table 8.9. DEA Mexican bank analysis, 13 banks, 1991. Output is total interest and non-interest income; inputs are total deposits and non-interest expense.

	CCR		BCC		GDEA					
Bank	θ	Class	θ	RTS	$\alpha = 10^3$ $(\boldsymbol{x}_o^T \nu = \boldsymbol{y}_o^T \mu)$	$\alpha = 15$	$\alpha = 10$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
(1) Banamex	0.531	NE	1.000	D	–181.32	0.00	0.00	0.00	0.00	0.00
(2) Bancomer	0.511	NE	1.000	D	–281.95	0.00	0.00	0.00	0.00	0.00
(3) Serfin	0.532	NE	1.000	D	–136.52	0.00	0.00	0.00	0.00	0.00
(4)Intermac	0.569	NE	0.908	–	–257.11	–717.26	0.00	0.00	0.00	0.00
(5) Cremi	0.704	NE	0.772	–	–282.58	–3,134.25	–1,957.76	0.00	0.00	0.00
(6) Bancreser	1.000	E	1.000	C	0.00	0.00	0.00	0.00	0.00	0.00
(7) MercNort	0.634	NE	0.638	–	–284.80	–4,371.50	–2,999.54	–385.14	–212.60	–42.31
(8) BCH	0.826	NE	0.828	–	–112.8	–1,481.79	–982.50	–99.34	–60.03	–15.61
(9) Confia	1.000	E	1.000	C	0.00	0.00	0.00	0.00	0.00	0.00
(10) Bancen	0.592	NE	0.612	–	–253.70	–1,621.77	–1,075.07	–50.54	0.00	0.00
(11) Promex	0.705	NE	0.715	–	–191.64	–2,262.34	–1,504.08	–74.49	0.00	0.00
(12) Banoro	0.535	NE	0.554	–	–295.19	–1,410.08	–934.00	–80.67	–5.37	0.00
(13) Banorie	0.937	NE	1.000	I	–73.42	0.00	0.00	0.00	0.00	0.00

C = constant returns to scale; D = decreasing returns to scale (RTS); E = efficient; I = increasing returns to scale; NE = not efficient.

has been shown that $GDEA_D$ can reveal domination relations among all decision-making units. It is expected from the results obtained in this study that GDEA will be useful for evaluating the efficiency of complex management systems in business, industry, and social problems.

References

Agrell, P.J., and Tind, J., 1998, An Extension of the DEA-MCDM Liaison for the Free Disposable Hull Model, Working Paper 3, 1–18, Department of Operations Research, University of Copenhagen, Copenhagen, Denmark.

Banker, R.D., Charnes, A., and Cooper, W.W., 1984, Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, **30**:1078–1092.

Belton, V., 1992, An integrating data envelopment analysis with multiple criteria decision analysis, in A. Goicoechea, L. Duckstein, and S. Zionts, eds, *Proceedings of the Ninth International Conference on Multiple Criteria Decision Making:Theory and Applications in Business, Industry and Commerce*, Springer–Verlag, Berlin, Germany, pp. 71–79.

Belton, V., and Vickers, S.P., 1993, Demystifying DEA: A visual interactive approach based on multiple criteria analysis, *Journal of Operational Research Society*, **44**:883–896.

Charnes, A., Cooper, W.W., Golany, B., Seiford, L.M., and Stutz, J., 1985, Foundations of data envelopment analysis for Preto-Koopmans efficient empirical production functions, *Journal of Econometrics*, **30**:91–107.

Charnes, A., Cooper,W.W., Huang, Z.M., and Sun, D.B., 1990, Polyhedral cone ratio DEA models with an illustrative application to large commercial banks, *Journal of Econometrics*, **46**:73–91.

Charnes, A., Cooper, W.W., and Rhodes, E., 1978, Measuring the efficiency of decision making units, *European Journal of Operational Research*, **2**:429–444.

Charnes, A., Cooper, W.W., and Rhodes, E., 1979, Measuring the efficiency of decision making units, *European Journal of Operational Research*, **3**:339.

Charnes, A., Cooper, W.W., Wei, Q.L., and Huang, Z.M., 1989, Cone ratio data envelopment analysis and multi-objective programming, *International Journal of Systems Science*, **20**:1099–1118.

Deprins, D., Simar, L., and Tulkens, H., 1984, Measuring labor-efficiency in post offices, in M. Marchand, P. Pestieu, and H. Tulkens, eds, *The Performance of Public Enterprises: Concepts and Measurements*, North Holland, Amsterdam, Netherlands, pp 247–263.

Farrell, M.J., 1957, The measurement of productive efficiency, *Journal of the Royal Statistical Society A*, **120**:253–281.

Golany, B., 1988, An interactive MOLP procedure for the extension of DEA to effectiveness analysis, *Journal of Operational Research Society*, **39**:725–734.

Halme, M., Joro, T., Korhonen, P., Salo, A., and Wallenius, J., 1999, A value efficiency approach to incorporating preference information in data envelopment analysis, *Management Science*, **45**:103–115.

Joro, T., Korhonen, P., and Wallenius, J., 1998, Structural comparison of data envelopment analysis and multiple objective linear programming, *Management Science*, **44**:962–970.

Korhonen, P., 1997, Searching the Efficient Frontier in Data Envelopment Analysis, Interim Report IR-97-79, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Sawaragi, Y., Nakayama, H., and Tanino, T., 1985, *Theory of Multiobjective Optimization*, Academic Press, London, UK, and New York, NY, USA.

Seiford, S.M., and Thrall, R.M., 1990, Recent developments in DEA: The mathematical programming approach to frontier analysis, *Journal of Econometrics*, **46**:7–38.

Stewart, T.J., 1996, Relationships between data envelopment analysis and multiple criteria decision analysis, *Journal of Operational Research Society*, **47**:654–665.

Taylor, W.M., Thompson, R.G., Thrall, R.M., and Dharmapala, P.S., 1997, DEA/AR efficiency and profitability of Mexican banks: A total income model, *European Journal of Operational Research*, **98**:346–363.

Thanassoulis, E., and Dyson, R.G., 1992, Estimating preferred target input-output levels using data envelopment analysis, *European Journal of Operational Research*, **56**:80–97.

Tulkens, H., 1993, On FDH efficiency: Some methodological issues and applications to retail banking, courts, and urban transit, *Journal of Productivity Analysis*, **4**:183–210.

Wei, Q.L., and Yu, G., 1997, Analyzing properties of k-cones in the generalized data envelopment analysis model, *Journal of Econometrics*, **80**:63–84.

Wierzbicki, A., 1980, The use of reference objectives in multiobjective optimization, in G. Fandel and T. Gal, eds, *Multiple Objective Decision Making, Theory and Application*, Springer-Verlag, New York, NY, USA.

Zhu, J., 1996, Data envelopment analysis with preference structure, *Journal of Operational Research Society*, **47**:136–150.

Yun, Y.B., Nakayama, H., and Tanino, T., 2000, On efficiency of data envelopment analysis, in Y.Y. Haimes and R.E. Steuer, eds, *Research and Practice in Multiple Criteria Decision Making: Proceedings of the XIVth International Conference on MCDM, Charlottesville, Virginia, USA, June 8–12, 1998*, Springer–Verlag, Heidelberg, Germany.

Chapter 9

Using Data Envelopment Analysis in Measuring Eco-Efficiency of Power Plants

Pekka Korhonen and Mikulas Luptacik

Abstract

In public discussion on environmental policy, the notion of eco-efficiency is often raised. The joint production of goods and undesirable outputs such as pollutants make it difficult to measure the overall performance of the firm, because those pollutants may not be freely disposable without costs. On the other hand, the lack of market prices for the undesirable outputs makes us unable to estimate harmful effects of pollutants in terms of costs. Some of the measurement and evaluation difficulties can be overcome when Data Envelopment Analysis (DEA) is employed as the efficiency measurement vehicle. This chapter considers two different approaches. In the first, we begin by decomposing the problem into two parts: 1) the problem of measuring technical efficiency (as the relation of the desirable outputs to the inputs) and 2) the problem of measuring so-called ecological efficiency (as the relation of desirable outputs to the undesirable outputs). Then we combine both these indicators. In the second approach, we treat the pollutants as the inputs in the sense that we wish to increase desirable outputs and reduce pollutants and inputs.

The approaches are applied to the problem of measuring the efficiency of 24 power plants in a European country.

Keywords: Technical efficiency, ecological efficiency, eco-efficiency, data envelopment analysis.

9.1 Introduction

One of the most intensively discussed concepts in the international political debate nowadays is the concept of *sustainability*. The great complexity of the notion of sustainable development requires new methodology for economic analysis and measurement of economic activities. A major issue concerns the question of how we get our economic accounting systems into a form where economic and ecological considerations are better taken care of than they are today. Increasing Gross National Product may cause harmful social and ecological effects. That's why we need new indicators to measure the economic performance of a firm or a national economy. On the occasion of the founder-meeting of the Austrian Business Council for Sustainable Development in July 1997 in Vienna, the Swiss entrepreneur Stephan Schmidheiny was quoted in the newspaper *Der Standard* (July 4, 1997) as saying: "There is no trade-off between economy and ecology." There must be a common denominator between the two, which he calls "eco-efficiency."

The main problem in developing eco-efficiency indicators is the lack of evaluations (like market prices) for wastes and emissions (in other words, for undesirable outputs). Some of these difficulties can be overcome when Data Envelopment Analysis (DEA) is used for efficiency measurement.

To our knowledge, the first paper using a non-parametric approach for multilateral productivity comparisons when some outputs are undesirable, is by Färe *et al*. (1989). For treating desirable and undesirable outputs asymmetrically, they use the enhanced hyperbolic output efficiency measure. This measure can be computed by solving a nonlinear programming problem: one takes a linear approximation of the nonlinear constraint. The methodology was applied to a sample of mills producing paper and pollutants. Other related papers are Färe *et al*. (1996) and Tyteca (1997). A comprehensive survey measuring the environmental performance of firms is provided by Tyteca (1996).

Golany *et al*. (1994) have considered the problem of measuring the efficiency of power plants using DEA, originally proposed by Charnes *et al*. (1978 and 1979) as a method for evaluating the Relative (Technical) Efficiency of Decision Making Units (DMUs), and essentially performing the same task. DEA also plays a key role in our approach.

The rest of this chapter is organized as follows. In Section 9.2 we discuss the different variants of DEA models which can be used for the estimation of eco-efficiency. It can be shown that the set of (strongly) efficient DMUs is the same for all models. In Section 9.3 we illustrate our methodology, using data for a sample of power plants in one European country. Concluding remarks are given in Section 9.4.

9.2 Theoretical Considerations

Assume we have n (homogeneous) DMUs, each consuming m inputs and producing p outputs. The outputs corresponding to indices 1, 2, . . . , k are desirable and the outputs corresponding to indices k+1, k+2, . . . , p are undesirable outputs. We prefer to produce desirable outputs as much as possible and to avoid producing undesirable outputs. Let $\mathbf{X} \in \mathbb{R}_+^{m\times n}$ and $\mathbf{Y} \in \mathbb{R}_+^{p\times n}$ be the matrices, consisting of nonnegative elements, containing the observed input and output measures for the DMUs. We decompose matrix $\mathbf{Y}$ into two parts: $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}^g \\ \mathbf{Y}^b \end{pmatrix}$, where a $k \times n$ -matrix $\mathbf{Y}^g$ stands for desirable outputs (“goods”) and a $(p-k)\times n$ matrix $\mathbf{Y}^b$ stands for undesirable outputs (“bads”). We further assume that there are no duplicated units in the data set. We denote by $\boldsymbol{x}_j$ (the j^{th} column of $\mathbf{X}$) the vector of inputs consumed by DMU$_j$, and by x_{ij} the quantity of input i consumed by DMU$_j$. A similar notation is used for outputs. Occasionally, we decompose the vector $\boldsymbol{y}_j$ into two parts: $\boldsymbol{y}_j = \begin{pmatrix} \boldsymbol{y}_j^g \\ \boldsymbol{y}_j^b \end{pmatrix}$, where the vectors $\boldsymbol{y}_j^g$ and $\boldsymbol{y}_j^b$ refer to the desirable and undesirable output-values of unit j. When it is not necessary to emphasize the different roles of inputs and (desirable/undesirable) outputs, we denote $\boldsymbol{u} = \begin{pmatrix} \boldsymbol{y}^g \\ -\boldsymbol{y}^b \\ -\boldsymbol{x} \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} \mathbf{Y}^g \\ -\mathbf{Y}^b \\ -\mathbf{X} \end{pmatrix}$.[1] Furthermore, we denote $\boldsymbol{1} = [1, \ \ldots, \ \ 1]^{\mathrm{T}}$ and refer by $\boldsymbol{e}_i$ to the i^{th} unit vector in $\mathbb{R}^{\mathrm{n}}$. We consider set T = $\{\boldsymbol{u} \mid \boldsymbol{u} = \mathbf{U}\,\lambda, \lambda \in \Lambda\}$, where $\Lambda = \{\lambda \mid \lambda \in \mathbb{R}_+^n$ and $\mathbf{A}\,\lambda \leq \boldsymbol{b}, \boldsymbol{e}_i \in \Lambda \,, i = 1, \ \ldots, \ n\}$.

Further consider matrix $\mathbf{A} \in \mathbb{R}^{1\times n}$, and vector $\boldsymbol{b} \in \mathbb{R}^1$, which are used to specify the feasible values of λ-variables.

In classical DEA, the measure of efficiency of a DMU is defined as a ratio of a weighted sum of (desirable) outputs to a weighted sum of inputs, subject to the condition that corresponding ratios for each DMU be less than or equal to one. The model chooses nonnegative weights for a DMU (whose performance is being evaluated) in a way that is most favorable for it. The *original model* proposed by

Charnes *et al*. (1978, 1979) for measuring the *technical efficiency* of unit '0', was as follows:

$$\max\ h_0 = \frac{\sum_{r=1}^{k} \mu_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \tag{9.1}$$

subject to:

$$\frac{\sum_{r=1}^{k} \mu_r y_{rj}}{\sum_{i=1}^{m} v_j x_{ij}} \leq 1,\ j = 1, 2, \ldots, n$$

$$\mu_r\ , v_i \geq \varepsilon,\ r = 1, 2,\ \ldots\ ,\ k;\ i = 1, 2,\ \ldots,\ m,$$
$$\varepsilon > 0 \text{ (“Non-Archimedean”).}$$

We refer to the unit under consideration by subscript '0' in the functional, but preserve its original subscript in the constraints. In Equation (9.1), only desirable outputs are used. The problem we will study in what follows is how to incorporate undesirable outputs into the model. There are at least two ways to approach the problem.

The first way is to decompose the problem into two parts and measure efficiency in two steps: first to measure a technical efficiency and then to measure another efficiency as a ratio of a weighted sum of (desirable) outputs to the weighted sum of (undesirable) outputs, called *ecological efficiency*. This leads to the following two models: The first, denoted *Model I* (Frontier Economics) is the standard DEA model [Equation (9.1)]. The second, for measuring the ecological efficiency (*Model II* – Deep Ecology) takes the form

$$\max\ g_0 = \frac{\sum_{r=1}^{k} \mu_r y_{r0}}{\sum_{s=k+1}^{p} \mu_s y_{s0}} \tag{9.2}$$

subject to:

$$\frac{\sum_{r=1}^{k} \mu_r y_{rj}}{\sum_{s=k+1}^{p} \mu_s y_{sj}} \leq 1,\ j = 1, 2,\ \ldots\ ,\ n$$

$$\mu_r \geq \varepsilon, r = 1,\ \ldots\ ,\ p.$$
$$\varepsilon > 0 \text{ (“Non-Archimedean”)}$$

The efficiency indicators of both models – in other words technical efficiency and ecological efficiency – are now the output variables for the new DEA model (with the inputs equal to 1), which yields the indicator for *eco-efficiency*.

The second approach is to build up the ratio, which simultaneously takes into account the (desirable) and (undesirable) outputs.

We will carry out the considerations by using the CCR model proposed by Charnes *et al*. (1978) ($\Lambda = \mathbb{R}^n_+$) , but the results can be generalized to other DEA models as well. We will review some approaches and show that the seemingly different models lead to similar results.

The first proposal is based on the idea of presenting all outputs as a weighted sum, but using negative weights for undesirable outputs. We call this "Model A" and give it as follows:

Model A:

$$\max\ h_{\mathrm{A}} = \frac{\sum_{r=1}^{k} \mu_r y_{r0} - \sum_{s=k+1}^{p} \mu_s y_{s0}}{\sum_{i=1}^{m} v_i x_{i0}} \tag{9.3}$$

subject to:

$$\frac{\sum_{r=1}^{k} \mu_r y_{rj} - \sum_{s=k+1}^{p} \mu_s y_{sj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1,\ j = 1, 2,\ \ldots,\ n$$

$$\mu_r,\ v_i \geq \varepsilon,\ r = 1, 2,\ \ldots,\ p;\ i = 1, 2,\ \ldots,\ m$$
$$\varepsilon > 0 \text{ (“Non-Archimedean”).}$$

Another possibility is to consider the undesirable outputs as inputs. This idea leads to the following approach, which is called Model B:

Model B:

$$\max\ h_{\mathrm{B}} = \frac{\sum_{r=1}^{k} \mu_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0} + \sum_{s=k+1}^{p} \mu_s y_{s0}} \tag{9.4}$$

subject to:

$$\frac{\sum_{r=1}^{k} \mu_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij} + \sum_{s=k+1}^{p} \mu_s y_{sj}} \leq 1,\ j = 1, 2,\ \ldots,\ n$$

$$\mu_r,\ v_i \geq \varepsilon,\ r = 1, 2,\ \ldots,\ p;\ i = 1, 2,\ \ldots,\ m$$
$$\varepsilon > 0 \text{ (“Non-Archimedean”).}$$

The third possibility is to consider the ratio of the weighted sum of the desirable outputs minus that of the inputs to that of the undesirable outputs. This idea leads to the following approach (Model C):
Model C:

$$\max\ h_{\mathrm{C}} = \frac{\sum_{r=1}^{k} \mu_r y_{r0} - \sum_{i=1}^{m} v_i x_{i0}}{\sum_{s=k+1}^{p} \mu_s y_{s0}} \tag{9.5}$$

subject to:

$$\frac{\sum_{r=1}^{k} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij}}{\sum_{s=k+1}^{p} \mu_s y_{sj}} \leq 1,\ j = 1, 2,\ \ldots,\ n$$

$$\mu_r,\ v_i \geq \varepsilon,\ r = 1, 2,\ \ldots,\ p;\ i = 1, 2,\ \ldots,\ m$$
$$\varepsilon > 0 \text{ (“Non-Archimedean”).}$$

We may also consider the reciprocal models of the models outlined above. The approach will lead to so-called “output”-oriented models, where the desirable outputs are controlled. As an example, we present Model D, which is a reciprocal model of Model B:

Model D:

$$\min\ h_{\mathrm{D}} = \frac{\sum_{s=k+1}^{p} \mu_s y_{s0} + \sum_{i=1}^{m} v_i x_{i0}}{\sum_{r=1}^{k} \mu_r y_{r0}} \tag{9.6}$$

subject to:

$$\frac{\sum_{s=k+1}^{p} \mu_s y_{sj} + \sum_{i=1}^{m} v_i x_{ij}}{\sum_{r=1}^{k} \mu_r y_{rj}} \leq 1,\ j = 1, 2,\ \ldots,\ n$$

$$\mu_r,\ v_i \geq \varepsilon,\ r = 1, 2,\ \ldots,\ p;\ i = 1, 2,\ \ldots,\ m$$
$$\varepsilon > 0 \text{ (“Non-Archimedean”).}$$

Using a standard technique see, e.g., Charnes *et al*., 1978, 1979) to transform the above fractional models [Equations (9.3)–(9.6)] into linear modes, we may use the following unified primal and dual presentation for all models. The presentation is called “Model G” (for more details, see Korhonen and Luptacik, 2000).

Note that the original primal formulation in Charnes *et al*. (1978) is sometimes in the DEA literature (see, e.g., Charnes *et al*., 1994) called the dual.

Model G:

General Model CCR Primal (CCR$_P$ - G)	General Model CCR Dual (CCR$_D$ - I)
$\max\ g_G = \sigma + \varepsilon\, \mathbf{1}^T(\mathbf{s}^b + \mathbf{s}^g + \mathbf{s}^-)$	$\min\ h_G = -\mu_g^T\, \mathbf{y}_0^g + \mu_b^T\, \mathbf{y}_0^b + v^T \mathbf{x}_0$
s.t. (10.7a)	s.t. (10.7b)
$\mathbf{Y}^g\lambda - \sigma\, \mathbf{w}^g - \mathbf{s}^g = \mathbf{y}_0^g$	$\mu_g^T\, \mathbf{w}^g + \mu_b^T\, \mathbf{w}^b + v^T\, \mathbf{w}^x = 1$
$\mathbf{Y}^b\lambda + \sigma\, \mathbf{w}^b + \mathbf{s}^b = \mathbf{y}_0^b$	$-\mu_g^T \mathbf{Y}^g + \mu_b^T \mathbf{Y}^b + v^T \mathbf{X} \geq \mathbf{0}$
$\mathbf{X}\lambda + \sigma\, \mathbf{w}^x + \mathbf{s}^- = \mathbf{x}_0$	$\mu_g,\ \mu_b,\ v \geq \varepsilon\, \mathbf{1}$
$\lambda, \mathbf{s}^-, \mathbf{s}^g, \mathbf{s}^b \geq \mathbf{0}$	$\varepsilon > 0$ ("Non-Archimedean")
$\varepsilon > 0$ ("Non-Archimedean")	

By choosing the components of vector $\mathbf{w} = \begin{pmatrix} \mathbf{w}^g \\ \mathbf{w}^b \\ \mathbf{w}^x \end{pmatrix}$ in a suitable way and modifying an objective function accordingly, we may introduce the corresponding presentations for each of the models A–D as shown in *Table 9.1* (see Korhonen and Luptacik, 2000).

Table 9.1. Required modifications of a general model.

Model type	$\mathbf{w}^g$	$\mathbf{w}^b$	$\mathbf{w}^x$	σ
A	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{x}_0$	$1-\theta$
B	$\mathbf{0}$	$\mathbf{y}_0^b$	$\mathbf{x}_0$	$1-\theta$
C	$\mathbf{0}$	$\mathbf{y}_0^b$	$\mathbf{0}$	$1-\theta$
D	$\mathbf{y}_0^g$	$\mathbf{0}$	$\mathbf{0}$	$-1+\theta$

Note that in the case of Models A–C, the value of the objective function $g_I = 1 - g_G$, I = A, B, C, and in case D: $g_D = g_G - 1$.

In data envelopment analysis, we are interested in the efficiency of the decision making units. Efficiency is defined as follows:

Definition 1. A point $\mathbf{u}^* = \mathbf{U}\lambda \in T$ is *efficient* iff (if and only if) there does not exist another $\mathbf{u} \in T$ such that $\mathbf{u} \geq \mathbf{u}^*$, and $\mathbf{u} \neq \mathbf{u}^*$.

The unit that is not efficient is called *inefficient*. However, if an inefficient unit is not an inferior point of T, we may call it *weakly efficient*. It is defined as follows.

Definition 2. A point $\mathbf{u}^* = \mathbf{U}\lambda \in T$ is *weakly efficient* iff there does not exist another $\mathbf{u} \in T$ such that $\mathbf{u} > \mathbf{u}^*$.

We may prove that eco-efficiency of a unit can be analyzed with Model G, and the result does not depend on vector $\mathbf{w} = \begin{pmatrix} \mathbf{w}^g \\ \mathbf{w}^b \\ \mathbf{w}^x \end{pmatrix}$ provided $\mathbf{w} \geq 0$, $\mathbf{w} \neq \mathbf{0}$ (Korhonen and Luptacik, 2000). The eco-efficient units are eco-efficient, no matter which model (A-D or G) is used, and eco-inefficient – but not weakly eco-efficient –

units can be diagnosed inefficient by using the value of σ at the optimum. For the weakly eco-efficient solution, not only the value of σ is sufficient.

9.3 Eco-Efficiency of Power Plants

In this section, we illustrate how we used our approach to evaluate the eco-efficiency and the emission reduction program of 24 power plants in a European country. The desirable output is electricity generation, with a minimum of 576,000 MW and a maximum of 2,160,000 MW. The total costs are considered as an input (min. US$ 1,345,448, max. US$ 13,014,761). The undesirable outputs or the pollutants are dust, NO_x, and SO_2. By the emission reduction program, the power plants reduced the emission quantities considerably. The emission levels are available before and after emission reduction. Before reduction, the emission quantities of dust (in tons/year) ranged between 574 and 14,097; after reduction they ranged between 175 and 1,418. The corresponding ranges for NO_x, in tons/year, were: before reduction [1,926; 5,509] and after reduction [963; 2,754]. For SO_2, in tons/year, levels were, before [1,401; 24,459], and after [1,401; 12,230].

Solving Models I [Equation (9.1)] and II [Equation (9.2)], we obtained the measures of the technical and ecological efficiency, respectively. Those measures provide the first indicators for the performance of power plants from the eco-efficiency point of view. The results are given in *Table 9.2*. The column denoted by “Technical efficiency" shows the results of Model I using total costs as an input and the electricity generation as an output. It is a very simple CCR model with only one efficient unit: namely, power plant 1, which is a small one with the lowest output level and the lowest total costs. Column 3 [“Ecological efficiency (before)”], presents the ecological efficiency *before* the emission reduction program. The results are obtained by solving Model II with electricity generation as the desirable output and with dust, NO_x, and SO_2 as pollutants or undesirable outputs. The ecologically efficient power plants are 1, 2, 4, 8, 13, and 14. The fourth column stands for the ecological efficiency *after* emission reduction. Only power plant 1 is technically and ecologically efficient – before emission reduction.

To get an indicator of the eco-efficiency, we took technical and ecological efficiency as output variables for the new DEA model with input equal to 1. In this way, the eco-efficiency is decomposed into technical and ecological efficiency. The eco-efficiency frontier before and after emission reduction is illustrated in *Figure 9.1* and *Figure 9.2*. These figures show that, before emission reduction, only power plant 1 is eco-efficient, and, afterwards, only power plants 1 and 2 are eco-efficient. The units 2, 4, 8, 13, and 14 (before emission reduction) and the units 4, 5, 8, 13, and 14 (after emission reduction) are only weakly efficient, because they are technically inefficient. Because all eco-inefficient units lie outside the eco-efficiency cone

Table 9.2. Technical and ecological efficiency analysis (CCR model).

Units	Technical efficiency	Ecological efficiency (before)	Ecological efficiency (after)
1	1.00	1.00	0.91
2	0.94	1.00	1.00
3	0.90	0.98	0.99
4	0.87	1.00	1.00
5	0.85	0.98	1.00
6	0.85	0.97	0.95
7	0.84	0.96	0.99
8	0.76	1.00	1.00
9	0.73	0.94	0.94
10	0.71	0.91	0.91
11	0.66	0.96	0.96
12	0.57	0.86	0.92
13	0.53	1.00	1.00
14	0.40	1.00	1.00
15	0.32	0.73	0.83
16	0.31	0.73	0.83
17	0.31	0.72	0.82
18	0.27	0.75	0.85
19	0.25	0.78	0.88
20	0.25	0.77	0.88
21	0.25	0.77	0.87
22	0.22	0.78	0.89
23	0.22	0.77	0.88
24	0.22	0.76	0.86

in both cases, the indicator of eco-efficiency is simply the better value of efficiency scores obtained from Models I and II. Thus the eco-efficient scores are the same as ecological efficiency scores in *Table 9.2,* except that Unit 1 is also eco-efficient after emission reduction.

To show the importance of the technical (ecological) efficiency in determining the eco-efficiency of units 1 and 2, we computed the ratio of weighted technical (ecological) efficiency to the virtual output (the weighted sum of technical and ecological efficiency). This is a useful indication of the importance of technical (ecological) efficiency in determining the eco-efficiency. In both power plants, the technical efficiency was given an importance of approximately 60% and the ecological efficiency of approximately 40% in determining the eco-efficiency. The strength of both units lies more in the area of technical efficiency, while the eco-inefficient units have a weakness primarily in the technical inefficiency. From the corresponding slack variables of the new DEA model, the potential eco-efficiency improve-

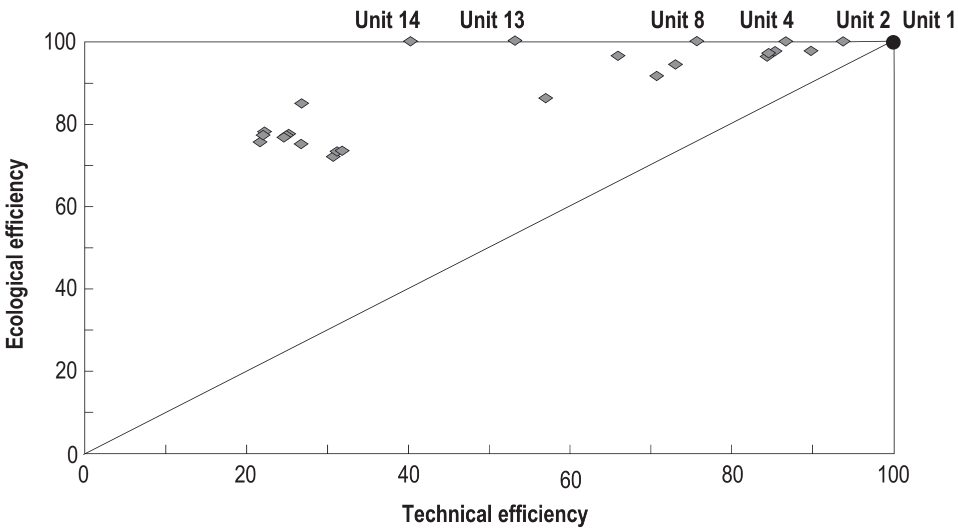


Figure 9.1. Eco-efficiency frontier before emission reduction.

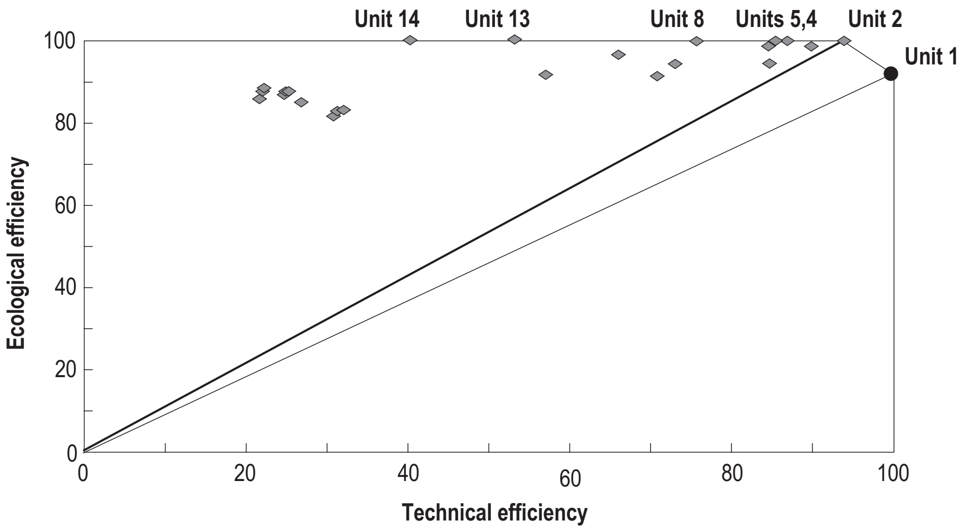


Figure 9.2. Eco-efficiency frontier after emission reduction.

ment with respect to technical and ecological efficiency, respectively, can be seen. It is obvious that the eco-efficiency after the reduction program is on average higher than before.

An alternative approach to analyzing eco-efficiency is to use models A, B, and C. *Table 9.3* shows the results of model B before and after emission reduction, both

Table 9.3. Eco-efficiency scores using the combined model B.

Units	Before emission reduction	After emission reduction
1	1.00	1.00
2	1.00	1.00
3	0.99	0.99
4	1.00	1.00
5	1.00	1.00
6	0.97	0.95
7	0.98	0.99
8	1.00	1.00
9	0.94	0.94
10	0.91	0.91
11	0.96	0.96
12	0.86	0.92
13	1.00	1.00
14	1.00	1.00
15	0.73	0.83
16	0.73	0.83
17	0.72	0.82
18	0.75	0.85
19	0.78	0.88
20	0.77	0.88
21	0.77	0.87
22	0.78	0.89
23	0.77	0.88
24	0.76	0.86

under the assumption of constant returns to scale. We will discuss in more detail the results of model B after emission reduction. A similar analysis can be done for the models A and C (before and after emission reduction) and for variable returns to scale.

The input variables in model B are total costs; the investment for emission reduction; and the emission of dust, NO_x, and SO_2 – all after the reduction program. The only output variable is electricity generation.

Comparing the results of model B (*Table 9.3*) with the eco-efficiency obtained as a composition of technical and ecological efficiency (*Table 9.2* and *Figure 9.2*) the tendency of the same results can be observed. Because DEA models yield the best possible results for every decision making unit, the eco-efficiency defined by model B cannot be lower than the eco-efficiency in *Figure 9.1*. For instance, the weakly eco-efficient power plants 4, 5, 8, and 13 from *Figure 9.1* (after) are eco-efficient according to model B. Units 1 and 2 are efficient in both cases. But model B provides a deeper insight into the causes of eco-inefficiency and shows the potential improvement to particular inputs and outputs. Nevertheless, decomposition of eco-efficiency into technical and ecological efficiency can be useful.

Computing the ratio of weighted inputs to the weighted sum of inputs we obtain an indication for the importance of particular inputs. For example, in units 2 and 5, an importance of 82% was given to abatement investment in determining their eco-efficiency. The investment in emission reduction was highly efficient. The strengths of unit 1 lie in its abatement investment (48%) and in the lower level of dust emission (47%). The abatement activity was oriented to reduction of NO_x only.

An interesting result is seen in power plant 8. The most important factor for the eco-efficiency of this unit is the low level of SO_2 emission (the lowest level of all power plants). This input was given an importance ranking of 99% in determining eco-efficiency.

The most important factors for power plant 13 are the abatement investment and the relatively low level of NO_x emission in comparison to the high level of output. Power plant 13 is the plant with the highest electricity generation. Power plant 14 is only weakly eco-efficient because of input inefficiency. The potential improvements lie in reducing abatement investment by 52% and in reducing total costs by 24%. Similar results can be found for the inefficient power plants 15–24. They have a weakness in technical efficiency and should primarily reduce their inputs.

9.4 Concluding Remarks

In this chapter we presented two approaches which can be used for the estimation of eco-efficiency. In the first approach, we measured the eco-efficiency in two steps: We estimated the technical efficiency and the so-called ecological efficiency separately. Then we took the results of both models as the output variables for the new DEA model (with the inputs equal to 1), which provides the indicator for eco-efficiency.

In the second approach, we formulated the different variants of DEA models, which simultaneously take into account the inputs, the pollutants or undesirable outputs, and the desirable outputs. It was shown that the efficient units are efficient, no matter which model variant is used. However, the efficiency scores may differ.

When one compares these two approaches, both tend to lead to the same results. However, the second approach provides a deeper insight into the causes of the eco-inefficiency and shows where potential improvements lie with respect to the particular inputs and outputs. The first approach yields the decomposition of eco-efficiency into technical efficiency and ecological efficiency.

As a topic for further research, we intend to introduce environmental standards into our models. In this way, we will be able to evaluate the impact of environmental policy on measures of efficiency, and to make multilateral productivity comparisons across the firms or particular industries in different countries.

Acknowledgment

This research was supported, in part, by grants from the Foundation of the Helsinki School of Economics and the Academy of Finland.

The topic of the chapter was one of the activities of the Decision Analysis and Support (DAS) Project at IIASA and research was started while the first author worked as the project leader of the DAS Project.

Note

[1] Because the results concerning $\boldsymbol{u}$ and $\mathbf{U}$ are also valid for $\begin{pmatrix} \boldsymbol{y}^g \\ -\boldsymbol{y}^b \\ -\boldsymbol{x} \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{Y}^g \\ -\boldsymbol{Y}^b \\ -\boldsymbol{X} \end{pmatrix}$, for simplicity, we often refer to $\boldsymbol{u}$ and $\mathbf{U}$, although we are factually interested in results concerning $\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix}$ and $\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix}$.

References

Banker, R.D., Charnes, A., and Cooper, W.W., 1984, Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, **30**:1078–1092.

Charnes, A., Cooper, W.W., and Rhodes, E., 1978, Measuring efficiency of decision making units, *European Journal of Operational Research*, **2**:429–444.

Charnes, A., Cooper, W.W., and Rhodes, E., 1979, Short communication: Measuring efficiency of decision making units, *European Journal of Operational Research*, **3**:339.

Färe, R., Grosskopf, S., Lovell, K., and Pasurka, C., 1989, Multilateral productivity comparisons when some outputs are undesirable: A nonparametric approach, *The Review of Economics and Statistics*, **71**(1).

Färe, R., Grosskopf, S., and Tyteca, D., 1996, An activity analysis model of the environmental performance of firms: Application to fossil-fuel-fired electric utilities, *Ecological Economics*, **18**:161–175.

Golany, B., Roll, Y., and Rybak, D., 1994, Measuring efficiency of power plants in Israel by data envelopment analysis, *IEEE Transactions on Engineering Management*, **41**(3):291–301.

Korhonen, P., and Luptacik, M., 2000, Eco-efficiency analysis of power plants: An extension of data envelopment analysis, Working Paper W–241, Helsinki School of Economics, Helsinki, Finland.

Tyteca, D., 1996, On the measurement of the environmental performance of firms: A literature review and a productive efficiency perspective, *Journal of Environmental Management*, **46**:281–308.

Tyteca, D., 1997, Linear programming models for the measurement of environmental performance of firms: Concepts and empirical results, *Journal of Productivity Analysis*, **8**:183–197.

Chapter 10

Time Series Prediction by Multi-Layered Neural Networks and Its Application to Prediction of Hourly Traffic Volume

Eiji Watanabe, Noboru Nakasako, and Yasuo Mitani

Abstract

This chapter discusses the time series prediction by using multi-layered neural networks. Here, a learning algorithm with forgetting is introduced and this learning algorithm is applied to the prediction problem of hourly traffic volume. From simulation results, we show that we can extract the characteristics of hourly traffic volume by using the learning algorithm with forgetting. Moreover, we show that it is very important to adjust the slope of the sigmoid function when predicting the hourly traffic volume.

Keywords: Time series prediction, hourly traffic volume, neural network, learning algorithm with forgetting, analysis of internal representations.

10.1 Introduction

In managing the natural environment, it is very important to construct the input-output model and/or the prediction model for natural resources. However, the changing properties of natural resources are generally complicated and often nonlinear.

It has been widely observed that the back propagation (BP) learning algorithm (Rumelhart *et al.*, 1986) is used for training multi-layered neural networks, and that they can approximate any continuous function within any precision (Funahashi, 1989). Neural networks have learning ability, parallel processing ability, and generalization ability. They have been widely applied to various engineering fields. Recently, neural networks have been actively applied to time series prediction (Gershenfeld and Weigend, 1994; Moriyama and Ishikawa, 1996). One of the advantages of neural networks is that they can automatically construct the mathematical model for time series data by learning. However, when we need good prediction performance, it is very important to determine the adequate structure of neural networks.

This chapter discusses time series prediction using multi-layered neural networks. We introduce a learning algorithm with forgetting (Ishikawa, 1994) and apply it to the prediction problem of hourly traffic volume. From simulation results, we show that we can extract the characteristics of hourly traffic volume by using the learning algorithm with forgetting. Moreover, we show that it is very important to adjust the slope of the sigmoid function for the prediction of hourly traffic volume.

10.2 Time Series Prediction by Auto-Regressive Model

The following model (Auto-Regressive or AR model) is widely used for modeling of time-series data because of its simplicity and effectiveness.

$$x_t + \sum_{i=1}^{p} a_i x_{t-i} = e_t, \;\; (t = 1, \cdots, N)\,, \tag{10.1}$$

where p and t denote the order of the AR model and time, respectively. e_t is assumed to be a white noise and has the following statistics:

$$E[e_t] = 0, \;\; E[e_{t_1} e_{t_2}] = \delta_{t_1 t_2} \sigma_e^2\,, \tag{10.2}$$

where δ denotes the delta function. When the number of time series data is large enough, the Yule-Walker method is often used to estimate the model parameters a_i.

$$\begin{cases} r_0 + \sum_{i=1}^{p} a_i r_i = \sigma_e^2 \\ r_\tau + \sum_{i=1}^{p} a_i r_{i-\tau} = 0 \ , \end{cases} \tag{10.3}$$

where the auto-correlation function r_τ is defined by $r_\tau = E[x_t x_{t-\tau}]$.

Moreover, the order p can be determined by AIC, the Akaike Information Criterion (Akaike, 1974):

$$AIC = N \log \sigma_e^2 + 2(p+1) \ . \tag{10.4}$$

In Equation 10.4, the order p that has the minimum AIC is selected as an optimal order.

10.3 Time Series Prediction by Multi-Layered Neural Networks

In the prediction of time series by neural networks, we consider the following model:

$$x_t + f(a_i, x_{t-i}) = e_t, \ \ (t = 1, \cdots, N) \ , \tag{10.5}$$

where $f(\cdot)$ denotes an unknown non-linear function. Then, a neural network is trained by the BP algorithm so as to minimize the error function:

$$E = \frac{1}{N} \sum_{t=1}^{N} (x_t - \hat{x}_t)^2 \ , \tag{10.6}$$

where $\hat{x}_t$ denotes the predicted value by a neural network.

Here, we should note the following problems in the usage of neural network models:

1. The number of input variables:
   It is very important to determine the adequate structure of neural networks, especially since the excessive number of hidden units may cause overfitting, which hinders generalization ability. Accordingly, many methods for the determination of the number of hidden units have been proposed (Reed, 1993).

2. The non-linearity of the system:
   Moreover, real data are usually governed by unknown non-linear functions. Although the non-linearity of neural networks should be adjusted so as to minimize the difference between the true system and the model, it is actually difficult to develop such an adjustment method. This chapter deals with such a difficulty by examining the relation between the slope of the sigmoid function and the prediction error in numerical experiments.

Ishikawa has proposed the learning algorithm with forgetting (Ishikawa, 1994) for the first problem. In this algorithm, the sum of weights is added to the error function E as follows:

$$J = E + \varepsilon' \sum_{i,j} |w_{ij}^{mm-1}| \,, \tag{10.7}$$

where w_{ij}^{mm-1} denotes the weight between the i-th and j-th units. ε denotes the amount of forgetting. The change of the weight given by (Ishikawa, 1994) is:

$$\Delta_J w_{ij}^{mm-1} = \Delta_E w_{ij}^{mm-1} - \varepsilon \mathrm{sgn}(w_{ij}^{mm-1}) \,, \tag{10.8}$$

with

$$\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \,,$$

where $\Delta_E w_{ij}^{mm-1}$ denotes the derivative of the error function E with respect to the weight w_{ij}^{mm-1}.

10.4 Prediction Results of Hourly Traffic Volume

In this chapter, we used hourly traffic volume observed by a vehicle detector for two months (from September 1, 1977 to October 30, 1977) in Fukuyama city, Japan. We used 300 points of data (from September 1 to 30) for modeling; others are used for prediction.

10.4.1 Effects by the number of input variables

First, the learning and prediction errors by both AR and neural network models are shown in *Figure 10.1*. In *Figure 10.1*, when the number of input variables exceeds

24, learning errors by the AR model decrease rapidly. This is due to the periodic characteristics of the hourly traffic volume. On the other hand, when the number of input variables is smaller than 24, learning errors by neural network models are smaller than those by the AR model. The neural network model is superior to the AR model with respect to the number of input variables.

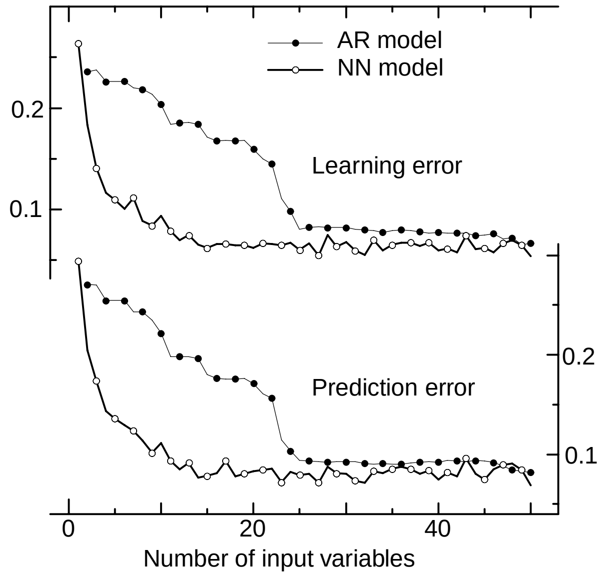


Figure 10.1. Learning and prediction errors for the number of input variables ($n_2 = 10$).

Figure 10.2 shows prediction results by both AR and neural network models. Here, the number of input variables is set at 13. These graphics show that a neural network model has better prediction ability than an AR model.

10.4.2 Relations between the internal representation and the prediction error

Learning and prediction results by the learning algorithm with forgetting (Ishikawa, 1994) are shown in *Figure 10.3*. If the amount ε of forgetting can be adequately selected, the prediction error can be reduced. It is difficult, however, to select an adequate range of the amount of forgetting. To address this difficulty, we introduce the following AIC$_\varepsilon$ criterion (Watanabe, 1996b):

$$AIC_\varepsilon = N \log \sigma_e^2 + 2(2^{H^2} + 2^{H^3}) , \tag{10.9}$$

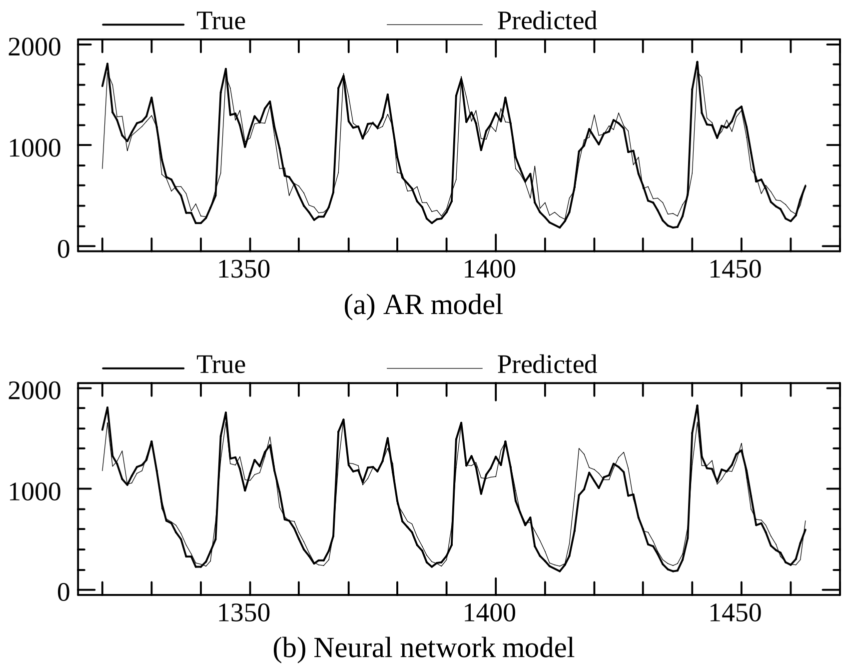


Figure 10.2. Prediction of hourly traffic volume.

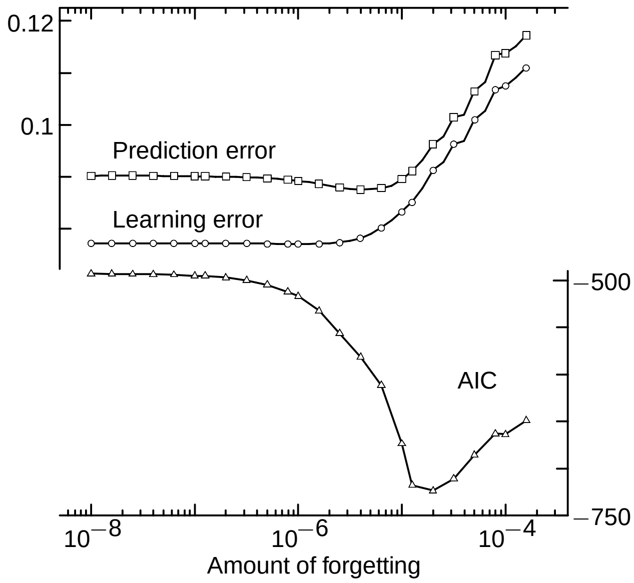
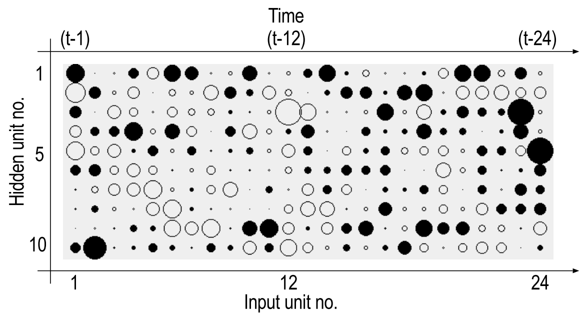
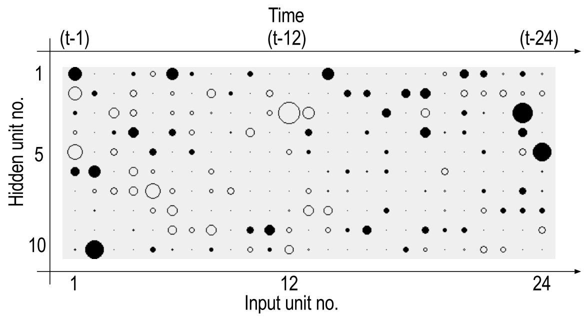


Figure 10.3. Learning results by the learning algorithm with forgetting.

where $H^m = |w_{ij}^{mm-1}| / \sum_{i,j} |w_{ij}^{mm-1}|$. The entropy H with respect to the weight was originally proposed by Ishikawa (Ishikawa and Uchida, 1992). When ε is near to 1×10^{-5}, both AIC$_\varepsilon$ and prediction error have the minimum, respectively. However, since the adequate range of the amount of forgetting is narrow, it is necessary to develop a modified method (Watanabe, 1996a), which can extend its range.



(a) By BP learning algorithm



(b) By learning algorithm with forgetting

Figure 10.4. Learning and prediction errors. (The size of each circle is proportional to the value of each weight. ○ denotes a weight with a negative value and • denotes a weight with a positive value. The line and row denote hidden and input units, respectively.)

In *Figure 10.4(a)*, weights by BP algorithm are, overall, distributed in input-hidden layers, and it is difficult to extract the rules obtained through learning. On the other hand, in *Figure 10.4(b)*, weights by the learning algorithm with forgetting are concentrated in specific units, and it is easier to extract the rules than it is with the BP algorithm. For example, it is confirmed that the weight between the third hidden unit and the 12th input unit has a negative value, and the weight between the third hidden unit and the 24th input unit has a positive value. Thus, the periodic characteristic of the hourly traffic volume can be automatically obtained by learning.

10.4.3 Relations between the slope of the sigmoid function and the prediction error

In modeling real data, it should be noticed that they are usually governed by unknown non-linear functions. Concretely, the non-linearity of the neural network can be represented as the slope r of the sigmoid function $f_r(x)$:

$$f_r(x) = \frac{1}{1 + e^{-rx}} \ . \tag{10.10}$$

Figure 10.5 shows learning and prediction errors for the change of the slope r of the sigmoid function. We can obtain a good prediction ability for neural networks, if we can adequately determine the slope r. In future work, we should consider control of the non-linearity of neural networks.

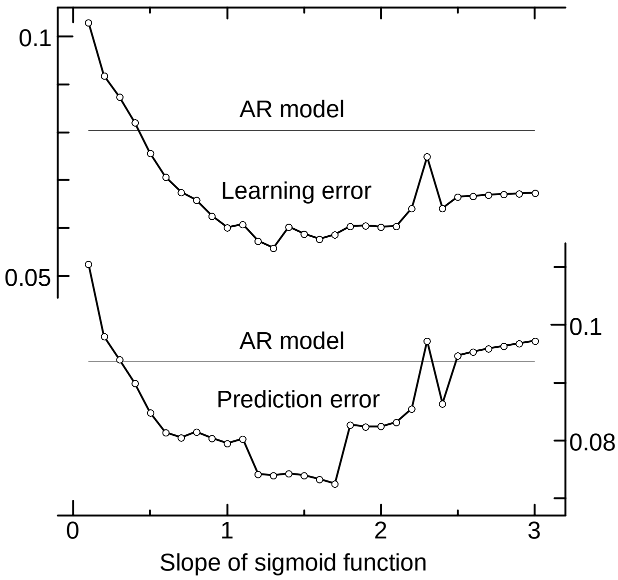


Figure 10.5. Learning results for the slope of the sigmoid function.

10.5 Conclusions

This chapter has discussed time series prediction by using multi-layered neural networks. The conclusions of this chapter can be summarized as follows:

1. From the viewpoint of the prediction error, neural network models are superior to AR models, when the number of input variables is relatively small.
2. The learning algorithm with forgetting (Ishikawa, 1994) makes it easier to extract the periodic characteristic of the traffic volume than does the original BP algorithm.
3. It has been shown that if we can determine the adequate slope of the sigmoid function, we can get a good prediction error by neural network models.

However, there are some problems to be solved in the future as follows:

1. The determination of the amount ε of forgetting and
2. The determination of the slope r of the sigmoid function.

Moreover, we would like to apply the neural network models for the construction of the input-output model and/or the prediction model for natural resources.

References

Akaike, H., 1974, A new look at the statistical model identification, *IEEE Trans. on Automatic Control*, **19**(6):716–723.

Funahashi, K., 1989, On the approximation realization of continuous mapping by neural networks, *Neural Networks*, **2**:183–192.

Gershenfeld, N.A., and Weigend, A.S., 1994, The future of time-series, *in* N. A. Gershenfeld and A. S. Weigend (eds), *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley, Reading, MA, USA, pp. 1–70.

Ishikawa, M., 1994, Structural learning in neural networks, in *Proceedings of the 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing*, Institute of Fuzzy Systems and Intelligent Control, Louisville, KY, USA, pp. 37–44.

Ishikawa, M., and Uchida, H., 1992, A structural learning of neural networks based on an entropy criterion, *Proceedings of IEEE/INNS Joint Conference on Neural Networks*, Vol. 2, Publishing House of Electronics Industry, Beijing, China, pp. 375–380.

Moriyama, T., and Ishikawa, M., 1996, Prediction of time series by a structural learning of neural networks, *Fuzzy Sets and Systems*, **82**:167–176.

Reed, R., 1993, Pruning algorithms–a survey, *IEEE Transactions on Neural Networks*, **4**(5):740–747.

Rumelhart, D.E., McClelland, J.L., and the PDP Research Group, 1986, *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA, USA.

Watanabe, E., 1996a, Selective learning algorithms with forgetting taking account of the balance between learning and forgetting, IEICE Technical Report NC95–171, IEICE, Tokyo, Japan, pp. 129–136.

Watanabe, E., 1996b, A study on relations between internal representation and generalization ability of multi layered neural networks for function approximation problems, IEICE Technical Report NC95–164, IEICE, Tokyo, Japan, pp. 77–84.

Chapter 11

Regression Analysis by a Mixture of Probabilistic Factor Analysis Models

Masahiro Tanaka and Masaaki Asada

Abstract

Regression analysis is designed to find the relation between a vector and a corresponding output. Recently, neural networks have often been used for this purpose. However, there are cases where the neural networks do not seem suitable. This chapter considers the application of nonlinear regression analysis, deploying a mixture of probabilistic factor analysis models. We will explain the usefulness of this kind of stochastic model and provide the estimation scheme. A joint probability density function of the input and output is the core tool, while the fundamental identification algorithm already has been proposed by Tipping and Bishop. We will use this model and the identification scheme, and will propose and discuss the technique for the regression problem.

Keywords: Regression analysis, mixture of probabilistic factor analysis, probability density function, multiple estimates, model determination.

11.1 Introduction

Regression analysis is designed to find the relation between the input vector and the output variable. In many cases, it is necessary for the estimator to be able to express the nonlinear relation. Multilayer neural networks (Rumelhart *et al*., 1986) are often applied for this purpose, but in some cases they are not particularly useful. For example, if the output is very noisy, a deterministic output does not mean much. Another case where a neural network is useless is when the output typically takes certain separate values stochastically.

The probability density function (PDF) describes the underlying distribution of the data itself, and hence it can be used in various analyses, including regression analysis. Moreover, it can to some degree overcome the problems mentioned above. Various marginal distributions can be obtained from the PDF directly, and the conditional expectations are also ready to be given.

A Gaussian model can be used as the first step for the PDF, but it yields only a linear function for the regression. A Gaussian mixture model may be used for a wide class of non-Gaussian models. However, when one observes the data locally, they often only exist in a subspace. Thus, numerical problems may arise. In such cases a lower dimensional model should be used. Thus the probabilistic factor analysis model (Tipping and Bishop, 1997) is a good candidate.[1] Since the data cannot be completely partitioned by the input cluster for expressing the PDF, we need an identification method for this model. Tipping and Bishop (1997) have proposed an identification scheme, but it lacks a method for determining the structure of the model.

This chapter consists of the following sections. In Section 11.2, the relation between the PDF and the regression analysis is explained. In Section 11.3, the Gaussian model and the Gaussian mixture models are introduced where the rank of the model is equal to the observation dimension. In Section 11.4, the probabilistic factor analysis (PFA) model and the mixture of PFA model are introduced, where the data may concentrate on subspaces locally. The identification algorithm by Tipping and Bishop will be shown. In Section 11.5, we propose determining the model structure by using AIC (Akaike Information Criterion). In Section 11.6, the regression analysis corresponding to the mixture of PFA model is shown. In Section 11.7, we demonstrate a result of the proposed regression analysis.

[1] Tipping and Bishop call this model the "principal component analysis model," but we will call it "probabilistic factor analysis model" or PFA model because of the form it takes in this chapter.

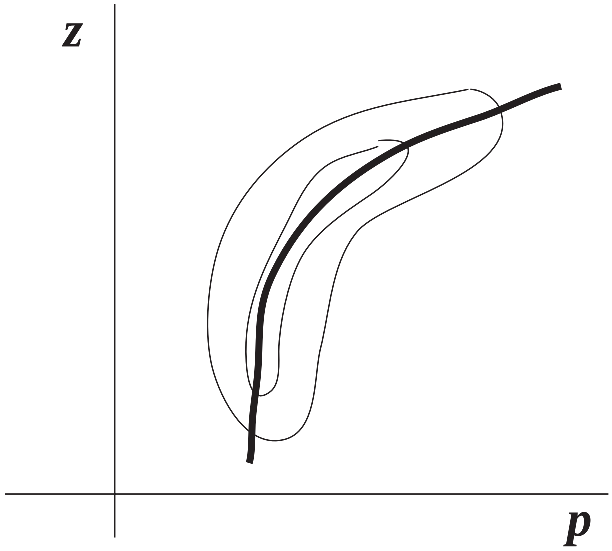


Figure 11.1. PDF and regression analysis (Case 1).

11.2 Probability Density Function and Regression Analysis

Regression analysis is a statistical method for estimating a variable z based on the vector consisting of the observed explanatory variables $\boldsymbol{p}$, i.e., the problem is to give the function

$$\hat{z} = f(\boldsymbol{p}) \ , \tag{11.1}$$

where the function $f(\cdot)$ is linear or nonlinear.

Figure 11.1 shows an example of the joint density input (horizontal line) and the output. The bold line indicates an appropriate output estimate for input values. Usually this is the conditional expectation $E[z|\boldsymbol{p}]$.

Figure 11.2 shows another example. If the joint distribution is like this, the output should be the bold line, not the mixed value that will lie where the density is very low.

It is well known (e.g., Anderson and Moore, 1979) that the minimum variance estimate of z based on the observation $\boldsymbol{p}$ is the conditional expectation

$$\hat{z}(\boldsymbol{p}) = E[z|\boldsymbol{p}] = \int zp(z|\boldsymbol{p})d\boldsymbol{p} \ , \tag{11.2}$$

whatever the PDF is. So, it is obviously important to estimate the conditional PDF $p(z|\boldsymbol{p})$, or equivalently, joint PDF $p(z, \boldsymbol{p})$ based on the training data set $\{(z(k), \boldsymbol{p}(k)); k = 1, \cdots, N\}$.

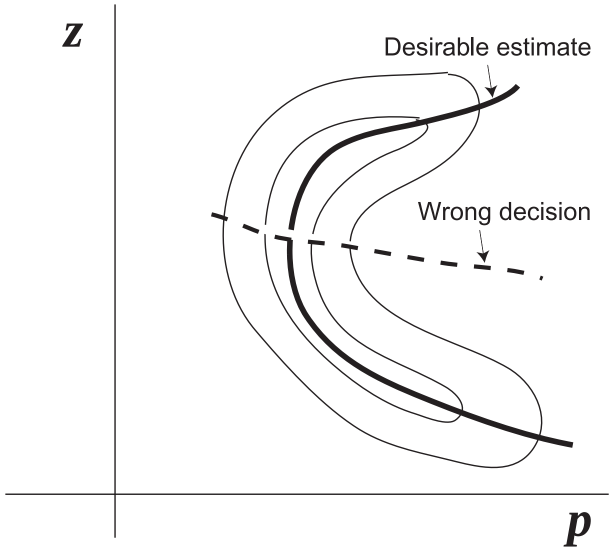


Figure 11.2. PDF and regression analysis (Case 2).

By the definition of the conditional probability, we have

$$p(z|\boldsymbol{p}) = \frac{p(z, \boldsymbol{p})}{p(\boldsymbol{p})} \ . \tag{11.3}$$

Thus, it is sufficient to estimate the joint PDF $p(z, \boldsymbol{p})$ to know $p(x|\boldsymbol{p})$. On the basis of this discussion, we will concentrate on estimating the joint PDF in the following sections.

11.3 Gaussian Mixture Model

11.3.1 Gaussian model

To simplify the notation, we use an augmented vector[2] $\boldsymbol{x} := [z, \boldsymbol{p}^T]^T$. A simple and most widely used PDF of a vector is a multivariate Gaussian function (e.g., Fukunaga, 1990) given by

$$p(\boldsymbol{x}) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right) \ , \tag{11.4}$$

where n is the dimension of $\boldsymbol{x}$, $\boldsymbol{\mu}$ is the mean vector, and Σ is the covariance matrix, which must be nonsingular.

[2]The superscript T denotes the matrix transposition.

Our problem is to estimate the parameters $\boldsymbol{\mu}$ and Σ based on the observation data $\{\boldsymbol{x}(1), \cdots, \boldsymbol{x}(N)\}$. If the maximum likelihood (ML) method is employed for this, the likelihood function should be defined as

$$L(\boldsymbol{\mu}, \Sigma) = p(\boldsymbol{x}(1), \cdots, \boldsymbol{x}(N)|\boldsymbol{\mu}, \Sigma)\ , \tag{11.5}$$

i.e.,

$$(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \text{arg}\ \max_{\boldsymbol{\mu},\Sigma}\ L(\boldsymbol{\mu}, \Sigma)\ . \tag{11.6}$$

In other words, the estimates are those which maximize the likelihood function. Within this context, if $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(N)$ are observed independently, it is easy to show that the ML estimates are given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}(k) \tag{11.7}$$

and

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N} (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T. \tag{11.8}$$

11.3.2 Gaussian mixture model

A functional form of the PDF that covers a wider class of distributions is the Gaussian mixture. It can be written as

$$p(\boldsymbol{x}) = \sum_i a^{(i)} p^{(i)}(\boldsymbol{x})\ , \tag{11.9}$$

where $a^{(i)}$ is a constant parameter with

$$\sum_i a^{(i)} = 1 \tag{11.10}$$

and $p^{(i)}(\boldsymbol{x})$ are Gaussian functions

$$p^{(i)}(\boldsymbol{x}) = (2\pi)^{-n/2}|\Sigma^{(i)}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^{(i)})^T(\Sigma^{(i)})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^{(i)})\right) \tag{11.11}$$

$$i = 1, \cdots, m\ ,$$

respectively. As the PDF is the sum of $a^{(i)}$-weighted distinct Gaussian kernel functions, we can easily interpret the meaning such that each observation instance $\boldsymbol{x}(k)$ comes from one of the Gaussian kernels $p^{(i)}(\boldsymbol{x})$ with probability $a^{(i)}$.

If the centers $\boldsymbol{\mu}^{(i)}$ are well-separated so that the clusters hardly overlap each other, we can identify the model by the following procedure:

1. Partition the data set according to the clusters.
2. Identify the parameters by using the partitioned subset of the data for each kernel.

However, it is usually impossible to partition the data into subsets corresponding to the Gaussian kernel completely.

The identification of the Gaussian mixture model can be obtained by using the Expectation-Maximization (EM) algorithm, where the "selection variable" is treated as the "missing data" in the EM algorithm terminology.

As we may notice, the Gaussian kernels may be degenerated, i.e., $\Sigma^{(i)}$ is singular for some i. Note that this may happen not only in the mixture model but also in the standard Gaussian model. For example, if z is a linear function free from observation noise, $\Sigma^{(i)}$ is singular.

Thus, we begin by considering how to cope with this problem for the Gaussian model.

11.4 Mixture of Probabilistic Factor Analysis Model

11.4.1 Probabilistic factor analysis model

In factor analysis, the following linear function is used:

$$\boldsymbol{x} = W\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e} \, , \tag{11.12}$$

where W is the factor loadings, $\boldsymbol{x}$ is the n-dimensional observation data, and $\boldsymbol{y}$ is an m-dimensional $(m < n)$ latent variable. The latent variable is a stochastic variable, and is assumed to obey the normal distribution

$$\boldsymbol{y} \sim N(0, I) \tag{11.13}$$

and

$$\boldsymbol{e} \sim N(0, \Psi) \, , \tag{11.14}$$

where Ψ is diagonal. Let us return to the model (11.12). Given this formulation, the observation vector is assumed to obey the normal distribution, with mean $\boldsymbol{\mu}$ and covariance $\Psi + WW^T$.

Using this model, the PDF of $\boldsymbol{x}$ can be written as

$$p(\boldsymbol{x}) = (2\pi)^{-n/2}|C|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) , \tag{11.15}$$

where

$$C = WW^T + \Psi . \tag{11.16}$$

It is possible to set

$$\Psi = \sigma^2 I \tag{11.17}$$

because of the existence of W. Tipping and Bishop (1997) have shown in this case that there is an explicit relationship between W and the principal axes of the covariance matrix of the observations S:

$$W = U_q(\Lambda_q - \sigma^2 I)^{1/2} R \tag{11.18}$$

$$S = \sum_{k=1}^{N} \frac{1}{N}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T , \tag{11.19}$$

where U_q is the q-column eigenvectors of S with corresponding eigenvalues in the diagonal matrix Λ_q and R is an arbitrary rotation matrix.

Given the observation $\boldsymbol{x}$, we can estimate the latent variable $\boldsymbol{y}$ by the minimum variance estimate (Anderson and Moore, 1979):

$$\hat{\boldsymbol{y}} = (WW^T + \sigma^2 I)^{-1} W^T(\boldsymbol{x}-\boldsymbol{\mu}) . \tag{11.20}$$

Coming back to our original problem, we can write that

$$\begin{bmatrix} z \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} W_z \\ W_p \end{bmatrix} \boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{e} . \tag{11.21}$$

Thus we have

$$\hat{\boldsymbol{y}} = W_p^T(W_pW_p^T + \sigma^2 I)^{-1}(\boldsymbol{p}-\boldsymbol{\mu}) \tag{11.22}$$

and

$$\hat{z} = W_zW_p^T(W_pW_p^T + \sigma^2 I)^{-1}(\boldsymbol{p}-\boldsymbol{\mu}) + \boldsymbol{\mu} . \tag{11.23}$$

11.4.2 Mixture of probabilistic factor analysis model

The model is just the extension of the PFA model to a mixture given by

$$p(\boldsymbol{x}) = \sum_i a^{(i)}(2\pi)^{-n/2}|C^{(i)}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^T(C^{(i)})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})\right) \tag{11.24}$$

where

$$C^{(i)} = W^{(i)}(W^{(i)})^T + (\sigma^{(i)})^2 I \; . \tag{11.25}$$

The identification algorithm for this mixture of PPCA model has been shown by Tipping and Bishop (1997). However, the dimension of each kernel (submodel) was assumed to be known *a priori*.

First we denote how the identification procedure of the PFA model is derived by using the EM algorithm (Tipping and Bishop, 1997).

The EM algorithm developed by Dempster *et al*. (1977) is an algorithm of ML estimate of the parameters where there are some missing data in the model. The missing data is a stochastic variable. The EM algorithm consists of two steps, and one must iterate them until the estimate converges. The E-step is to take expectation of the complete likelihood, based on the observation and the model parameters that have been obtained up until now. The M-step is to update the parameters so that the expectation of the complete likelihood is to be maximized.

11.4.3 Identification of PFA model

For the PFA model, the latent variable can be treated as the missing data. The complete data log-likelihood is given by

$$L_c = \sum_k \log p(\boldsymbol{x}(k), \boldsymbol{y}(k)) \; . \tag{11.26}$$

Identification Algorithm

Initialization Let W and σ^2 be some random values.

Iteration Iterate the following steps until the parameters converge.

- **E-step** Take the conditional expectation of L_c where the distribution used for this purpose is $p(\boldsymbol{y}(k)|\boldsymbol{x}(k), \boldsymbol{\mu}, W, \sigma^2)$.
- **M-step** Update $\boldsymbol{\mu}$, W and σ^2 so that the conditional expectation of the complete likelihood of the E-step is maximized.

11.4.4 Identification of mixture of PFA model

The likelihood function we actually want to maximize is

$$L = \sum_k \log \left(\sum_i a^{(i)} p^{(i)}(\boldsymbol{x}(k) | \boldsymbol{\mu}^{(i)}, W^{(i)}, (\sigma^{(i)})^2) \right) . \tag{11.27}$$

For the identification of the mixture of PFA model, the complete likelihood function can be defined as

$$L_c = \sum_k \sum_i I_i(k) \log \left(a^{(i)} p(\boldsymbol{x}(k), \boldsymbol{y}^{(i)}(k) \right) , \tag{11.28}$$

where $I_i(k)$ is a stochastic binary variable and $\boldsymbol{y}^{(i)}(k)$ is the latent variable vector of the i-th sub-model. The unknown values to be identified are $a^{(i)}$, $\boldsymbol{\mu}^{(i)}$, $W^{(i)}$ and $(\sigma^{(i)})^2$. However, as is pointed out by Tipping and Bishop (1997), the update equations of $\boldsymbol{\mu}^{(i)}$ and $W^{(i)}$ need to use the updated values of $W^{(i)}$ and $\boldsymbol{\mu}^{(i)}$, respectively; thus they cannot be computed by evaluating only once. This leads to a huge computational load. Tipping and Bishop (1997) have developed a two-stage EM procedure.

The first stage is to update $a^{(i)}$ and $\boldsymbol{\mu}^{(i)}$ by other fixed parameters. The second stage is to update $W^{(i)}$ and $(\sigma^{(i)})^2$ where yet others are fixed. This corresponds to GEM (generalized EM) and it is guaranteed that it converges to a local maximum of the likelihood function (Dempster *et al*., 1977).

11.5 Model Determination by Using AIC

It is necessary to determine the dimensions of the latent variables. The best model can be determined which minimizes a criterion like AIC (Akaike Information Criterion). AIC is defined as

$$AIC = -2L + 2P , \tag{11.29}$$

where L is the likelihood function and P is the number of free parameters. The problem is defined as to minimize this criterion.

If the number of the kernels is small and the observation dimension is also small, it is possible to calculate it by an exhaustive search. However, as these numbers grow, the search space expands exponentially, hence some heuristic method may be necessary. Meta-heuristics such as the genetic algorithm also may be useful in some cases. Preliminary experiments have been done by Tanaka (2000).

11.6 Regression Analysis Based on Mixture of PFA

The estimate of z based on the observation $\boldsymbol{p}$ is given by

$$\begin{aligned}\hat{z} &= E[z|\boldsymbol{p}] = \int zp(z|\boldsymbol{p})dz \\ &= \sum_i E[z|\boldsymbol{p}, i]P(i|\boldsymbol{p}) \,, \end{aligned} \tag{11.30}$$

where the conditional expectation of the kernel i given the input is

$$E[z|\boldsymbol{p}, i] = \sum_i W_z^{(i)}(W_p^{(i)})^T \left(W_p^{(i)}(W_p^{(i)})^T + (\sigma^{(i)})^2 I\right)^{-1} (\boldsymbol{p} - \boldsymbol{\mu}^{(i)}) + \boldsymbol{\mu}^{(i)} \;; \tag{11.31}$$

the *a posteriori* probability of the event occurrence from the kernel i is

$$P(i|\boldsymbol{p}) = \frac{p^{(i)}(\boldsymbol{p})P^{(i)}}{\sum_i p^{(i)}(\boldsymbol{p})P^{(i)}} \;; \tag{11.32}$$

further where the stochastic density function of the input $\boldsymbol{p}$ for the kernel i is

$$p^{(i)}(\boldsymbol{p}) = \sum_i (2\pi)^{-n/2}|C^{(i)}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{p} - \boldsymbol{\mu}_p^{(i)})^T(C_p^{(i)})^{-1}(\boldsymbol{p} - \boldsymbol{\mu}_p^{(i)})\right) \tag{11.33}$$

$$C_p^{(i)} = W_p^{(i)}(W_p^{(i)})^T + (\sigma^{(i)})^2 I \;; \tag{11.34}$$

and $P^{(i)}$ is the *a priori* probability of the kernel i. Note that this is quite similar to the form of a certain kind of fuzzy model (e.g., Tanaka, 1998).

Figure 11.3 shows the structure of the model. From each model, an output is computed based on the input value. The estimator for each model is a linear one [Equation (11.33)]. The mixing weight is a function of the input $\boldsymbol{p}$, and hence the final output is a nonlinear function of the input.

In certain cases, it is better not to mix the estimate of all the kernels. As was mentioned in Section 11.1, we may sometimes encounter a case where the output sorts into separate groups. Taking the average value of the estimates in this situation may severely degrade the output estimate. Such circumstances can occur when an important attribute is not included in the model. The output appears to be taking quite distinct values stochastically, and it is obviously better to propose the estimate as a set. To do this, it is necessary to group the kernels, and within the group the outputs are mixed. This could be done by checking the sum of the *a posteriori*

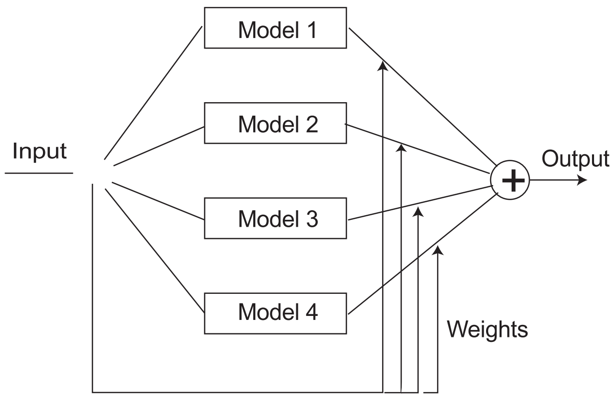


Figure 11.3. Structure of the model.

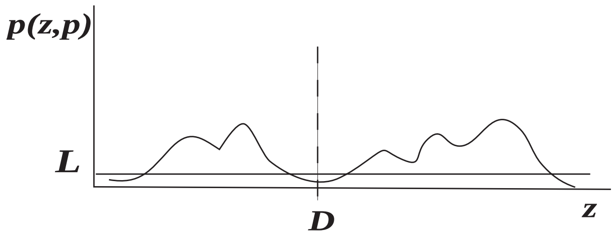


Figure 11.4. How to find the disconnection of the output.

probabilities. If extremely low values are revealed, we can judge that the output is disconnected there.

Figure 11.4 shows an idea for doing this. The horizontal axis is the output value, and the vertical axis is the joint PDF of the input and output. Note that the input is fixed to the value of our interest. By scanning the joint PDF along the output value, we may find the point where the joint PDF takes a very low value (Point D in *Figure 11.4*). If we take this as a disjoint point, the mixture of the output is done in the right part and the left part, separately.

11.7 Numerical Example

We have done some preliminary experiments for the algorithm proposed in this chapter. *Figure 11.5* shows the AIC values for the data where the number of kernels is 5; 200 points were generated from each kernel; and the dimension of $\boldsymbol{y}$ was 5 for all the kernels except one. In that kernel, data were generated where the dimension of the latent variable $\boldsymbol{y}$ was varied to 1 to 5 in each of the 5 experiments. For each experiment, the AIC values are plotted in *Figure 11.5* where the dimension of that

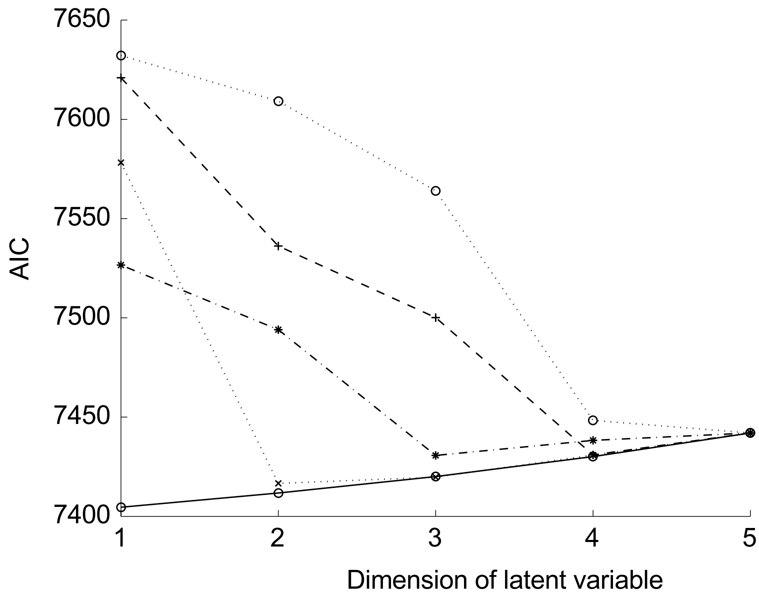


Figure 11.5. Dimension and AIC.

variable was searched by increasing from 1 to 5. The figure shows that the correct dimension yields the minimum AIC values for all the experiments.

For the regression problem, we generated 500 points of 3-dimensional data by using a parameter θ as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sin(\theta) \\ \cos(\theta) \\ \theta \end{bmatrix} + e \ , \tag{11.35}$$

where θ varied from 0 to 6π.

In this experiment, x and y are used as the input and z is used as the output. z is obviously a multimodal function of x, y. *Figure 11.6* shows the data points and the center of the kernels; 350 points were used for the identification and 150 points were used for the evaluation of the model.

Table 11.1 and *Table 11.2* show the *a posteriori* probabilities of the kernels and the estimates of z for $(x, y) = (0, -1)$. The final estimate of z is 9.648.

11.7.1 A simple mixture model and the regression

We generated data based on

$$\boldsymbol{x} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \boldsymbol{y} + \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \boldsymbol{e} \ \text{ w.p. } 0.3 \tag{11.36}$$

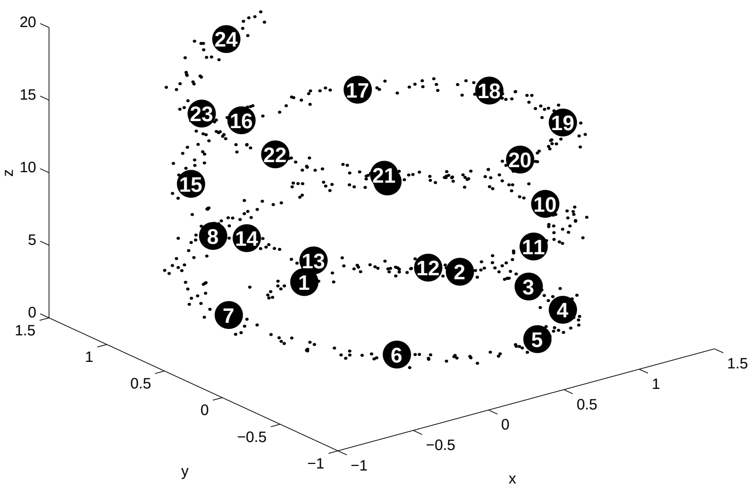


Figure 11.6. Centers of the generated kernels.

and

$$\boldsymbol{x} = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix} \boldsymbol{y} + \begin{bmatrix} 2 \\ 5 \end{bmatrix} + \boldsymbol{e} \text{ w.p. } 0.7 , \tag{11.37}$$

where all the variables are 2-dimensional.

In this experiment, x_1, the first element of the observation vector, is the output z, and x_2, the second element of the observation vector, is the input p. In this experiment, we assume that all the parameters are already known, and our problem is to get the estimate of the output z based on the input p.

Figure 11.7 plots the data $\boldsymbol{x}$.

Figure 11.8 shows the estimate of the output z given the input p. To clarify the result shown on the figure, we estimated about 20% of all the data. The symbol $\circ$ denotes the output estimate where the probability $P(i|p)$ is more than 0.9. The symbol $\times$ denotes the output estimate where the probability $P(i|p)$ is less than 0.1. The symbol $\triangle$ denotes those where the probability is more than 0.1 and less than 0.9. This shows that our estimation scheme yields an appropriate result for the estimate of multiple outputs with probability.

Table 11.1. *A posteriori* probabilities and the estimates for the kernels.

Kernel #	Probability	Estimate
1	0	–0.4024
2	0	0.8930
3	0	5.1589
4	0	2.8596
5	0.139	3.3523
6	0.105	3.2693
7	0	3.5590
8	0	5.4066
9	0	6.9769
10	0	9.2003
11	0.333	9.6141
12	0.204	9.5174
13	0	9.1545
14	0	9.6840
15	0	9.4458
16	0	12.4068
17	0	12.3739
18	0	15.2860
19	0	14.4008
20	0.216	16.9626
21	0.002	15.8983
22	0	15.9508
23	0	16.9339
24	0	18.5643

Table 11.2. A set of estimates for the same input.

Probability	Estimate
0.245	3.317
0.537	9.577
0.218	16.953

11.8 Conclusions

In this chapter, the regression scheme has been shown for the PFA model. This model can treat a wide class of data distribution where output may take multiple distinct values. Such a case can happen when an important explanatory variable is not included in the model, as is possible with environmental data. In this chapter, the model determination issue has not been extensively treated in the experiment. Further experiment is needed for validating this method for actual problems.

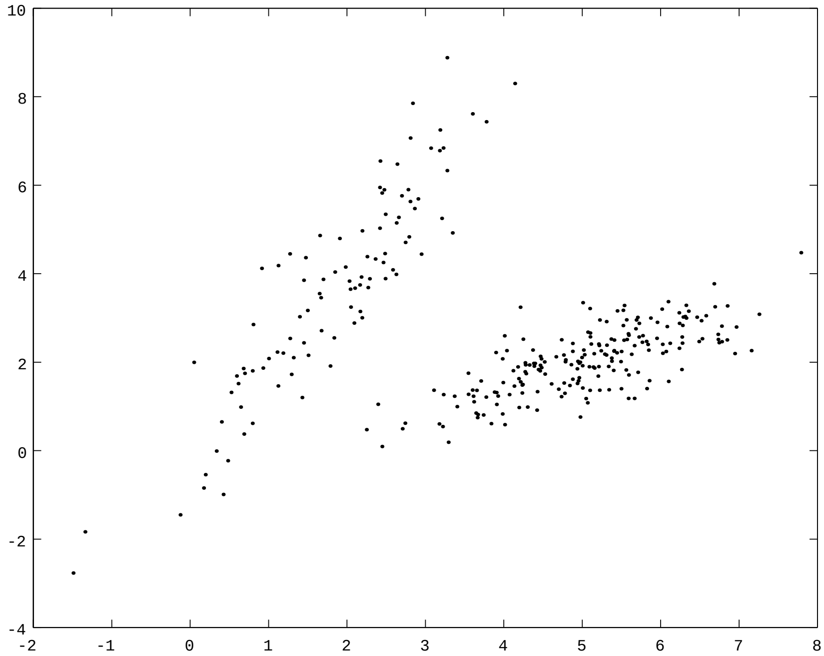


Figure 11.7. Observation data.

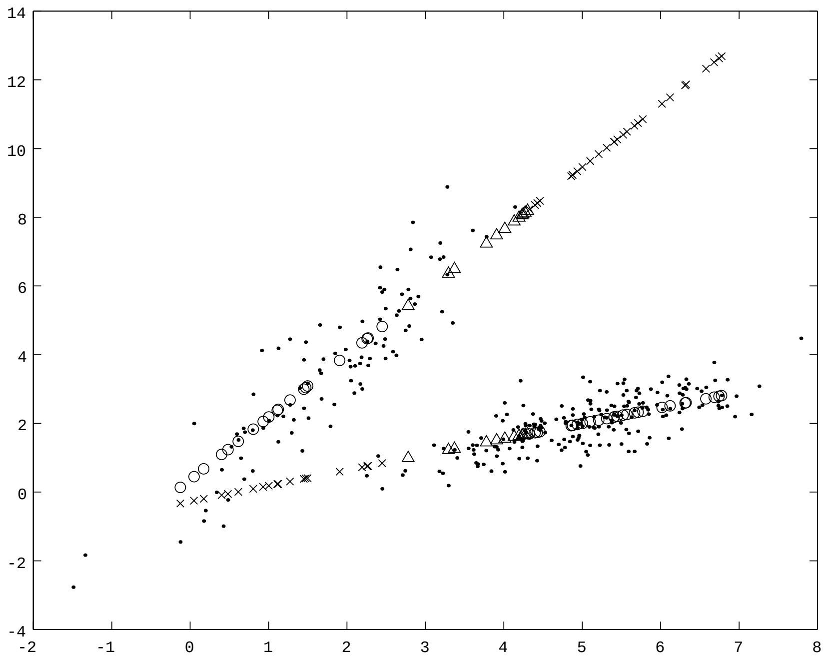


Figure 11.8. Observation data and the estimates of the outputs.

Acknowledgment

Second author was engaged in this work when he was a student at Okayama University.

References

Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **B39**(1):1–38.

Anderson, B.D.O., and Moore, J.B., 1979, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, USA.

Fukunaga, K., 1990, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, New York, NY, USA and London, UK.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J., 1986, Learning internal representation by error propagation, in D.E. Rumelhart *et al*., eds, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.1: Foundations*, The MIT Press, Cambridge, MA, USA, pp. 321–362.

Tanaka, M., 2000, Modeling of mixtures of principal component analysis model with genetic algorithm, Proceedings of the 31st ISCIE International Symposium on Stochastic Systems Theory and Its Applications, Institute for System Control and Information Engineers, Kyoto, Japan, pp. 157–162.

Tanaka, M., 1998, *An Introduction to Softcomputing*, Science and Technology Publishing Company, Inc., Tokyo, Japan (in Japanese).

Tipping, M.E., and Bishop, C.M., 1997, Mixtures of Probabilistic Principal Component Analysers, Technical Report NCRG/97/003/, Aston University, Birmingham, UK, 29 pages.

Part III: Land Use

Chapter 12

Land-Use Change in China: Combining Geographic Information System and Input-Output Analysis

Klaus Hubacek and Laixiang Sun

Abstract

Land availability is of crucial importance for China's development in the 21st century. Economic growth, urbanization, changes in life styles, and population growth will influence both the demand for and the supply of land. In this chapter, an input-output model expanded by a set of land categories was developed to synthesize various scenarios of change in the economy and society, and to evaluate their impact on land-use changes in China. The spatial aspect is provided by a number of fairly large and detailed geographical databases on China, including biophysical attributes of land and demographic data at the county level, that have been implemented in a geographical information system. The scenario analysis is conducted at both the national and regional levels for a time horizon of over 30 years. The analysis shows how different development paths will influence the available land base as well as the inter-regional and international trade flows of primary products for China in the coming decades.

Keywords: Input-output analysis, geographical information system (GIS), scenario analysis, China, land use change, agro-ecological zones model, hybrid models.

12.1 Introduction

Land availability is crucial for China's food security and economic development. Although China has a total area of some 960 million hectares (ha), which is the third largest in the world, only about 14.8% are cultivated with field crops and horticultural products. Primary farmland is located mainly in the same geographic areas where population and major economic activities have been concentrated. About one billion people (out of China's population of 1.3 billion) are concentrated in less than one third of the land area. The eastern region (Yangtze Delta), Sichuan, and the urban conglomerations along the eastern coast are the main population centers. These coastal areas are also the ones experiencing the highest growth rates in the economy. In several eastern provinces, settlement areas already cover more than 10% of the total land and are expanding further. Cropland areas are shrinking due to both urban sprawl and the growing land requirements of villages, rural industries, and infrastructure. China lost some 980,000 ha of cultivated land to construction activities between 1988 and 1995 (Heilig, 1999).

China's food security can be threatened by loss of cultivated land due to disasters, water and wind erosion, as well as chemical and physical deterioration. Agricultural over-exploitation and industrial pollution also exacerbate these degradation problems. Even though there are some controversial arguments about food demand and supply in China for the next 30 years (Brown, 1995; Chen *et al*., 1996; Huang and Rozelle, 1997), there is agreement that arable land loss and land degradation are undermining China's food production capacity (e.g., Gardner, 1996; Rozelle and Huang, 1997). In the case of forestland and grassland, over-exploitation and degradation might be even more severe (Fischer *et al*., 1996; Liu, 1998; Richardson, 1990).

Another trend in changing land use is agricultural restructuring: the transformation of China's cropland into horticultural land and fishponds. This shift is owing to changes in consumer demand as well as institutional and supply-side factors. It has become much more profitable for Chinese farmers to grow vegetables and fruit and sell these for market prices rather than to produce rice or wheat, which are still regulated by the state's procurement system. The conversion of cropland into fishponds and horticultural lands following the market-driven restructuring requirements of the agricultural sector would actually increase food security. The conversion of cropland into forest and grassland, arising from the requirement of conserving soil resources and the environment, is also desirable from a long-term perspective.

These changes in China's land-use patterns reflect changes in the country's institutional framework, economy, and society. China has been changing from a command-based economy to a market-based one, resulting in annual gross domestic product (GDP) growth rates of 9.8% between 1978 and 1998. Increased income and migration from rural to urban areas have caused changes in lifestyles. These changes are compounded by China's large population.

In this chapter, we develop a number of scenarios around the major driving forces in China's economy and society: technical change, income growth, changing patterns in consumption and production, urbanization, and population growth. The basic reference year is 1992 and the year for scenario analysis is 2025.

In order to assess how changes in the economy and society affect future land use, it is necessary to combine biophysical, economic, and societal data. A consistent theoretical framework is very important for such investigations. In this chapter, we employ a structural economics framework in which scenarios about possible future stages of society, economy, and the environment are embedded. The core of our framework is an input-output (I-O) model with strong biophysical linkages to allow the assessment of land-use change.

Socioeconomic changes are linked to different types of land via an explicit representation of land requirement coefficients associated with specific economic activities. In this way, land is treated as explicit factor input. The strong biophysical linkages are mainly manifested in the derivation of regional differences of the land requirement coefficients and the typical I-O technical coefficients. In other words, while we can stylize certain technological development trends at the national level based on a literature survey, their regionalization is not straightforward. Therefore we create the regionalized linkages based on the Agro-Ecological Zone (AEZ) assessment within a Geographical Information System (GIS). In addition, the AEZ assessment is also used to derive the future land suitability in each region.

In many studies dealing with similar questions, the focus has been either on a small region or on all of China. The small-region models might deliver excellent results for the region concerned, but they are unable to deal with the interplay across regions and do not allow any predictions for the national level. Studies focusing on the national level usually lack the capability to tackle regional differences and the interaction among regions. Typically, population densities, soil and climate conditions, and economic development are significantly different across regions in a large developing country like China. China can be perceived as a group of co-evolving, disparate economies rather than as a homogeneous entity. On one hand, China has fast-developing urban growth centers in the coastal areas and, on the other hand, backward rural areas that are each associated with distinct income, lifestyle, and expenditure patterns. Differing regional growth paths in the past might also have considerable path-dependent effects in the future and influence the future flow of regional migration due to labor demand of growth centers.

In this chapter, we build our model from the ground up and develop seven regional models and then a national one for China. We specify various development paths for different regions using data and information available at both the regional and national levels. The combination of and communication between regional and national models enable us to investigate how the constraints of land availability in each region might affect the inter-regional trade flow of land-based products. This relationship further allows us to evaluate the degrees of land scarcity at both regional and national levels and the extent of the necessary land-productivity improvement that is desired for keeping the land requirement feasible in the future. As far as we know, our model is among the very first to set up inter-related regional I-O models for China with strong biophysical linkages explicitly focusing on land use change.

The output of this chapter will provide not only a primary assessment of land-use feasibility with respect to selected scenarios that may represent possible directions of the Chinese economy and society in the future, but also an initialization for the dynamic welfare optimum model of IIASA's Land-Use Change (IIASA-LUC) Project (see also Chapter 16 in this book). The dynamic welfare optimum model of IIASA-LUC intends to establish a more integrated assessment of the spatial and intertemporal interactions among various socioeconomic and biogeophysical factors that drive land-use and land-cover change. It aims also to trace the possible adaptive behaviors of economic actors and the resulting consequences under the condition of increasing scarcity of land resources (Sun, 2000).

12.2 Linking the Basic Input-Output Model with Bio-Physical Data

The core of our approach is an I-O model expanded by a set of land categories. The rationale for extending the standard I-O framework to estimate land-use change can be summarized as follows. In order for the final demand of a given sector to expand, the output of other sectors must expand as well, corresponding to the input requirements of the given sector. Since all economic activities consume space, in the long run, in order to achieve significant increases in output, there must be increases or changes in land use or land productivity.

In the standard version, changes in the exogenously given vector of final demand (Δy) are driving the economy via a matrix of output multipliers, the Leontief inverse, $(I - A)^{-1}$, resulting in changes in sectoral output (Δx):

$$(I - A)^{-1}\Delta y = \Delta x \ . \tag{12.1}$$

In order to link land-use changes in economic sectors to those in land categories (such as cultivated land, grassland, forestland, etc.), the vector representing changes in output (Δx) is pre-multiplied by a diagonal land requirement coefficient matrix ($\widehat{C}$) and a land distribution matrix (R). The land distribution matrix R gives the mapping relationship between land uses in economic sectors and the natural categories of land, and the attributes in R are the shares of the former in the latter (see *Table 12.3*). The land requirement coefficient vector (c_j) is defined as the ratio of total land use in each sector (L_j) over total sectoral output (x_j), representing land use in ha per one million Yuan of output of sector j. The land requirement coefficient (c_j) is equivalent to the inverse of sectoral land productivity (p_j), which represents the output in Yuan produced on one ha of land. The future land use (L^{2025}) is the sum of the present land uses (L^{1992}) and the changes in land use (ΔL), triggered by the changes in output (Δx) based on the scenarios:

$$\Delta L = R\widehat{C}\Delta x \ , \tag{12.2a}$$

$$L^{2025} = L^{1992} + \Delta L \ . \tag{12.2b}$$

Standard I-O models usually assume that the economy instantaneously (that is, within the observed time period, usually a year) adjusts to shifts in spending patterns. All production activities are assumed to be endogenous and demand-driven, due to the assumed excess capacity throughout the economy. Supply is assumed to be perfectly elastic in all sectors, and an increase in demand is sufficient to stimulate increases in output and incomes. However, it is clear that some sectors will not automatically expand or shrink their land requirements in direct proportion to output changes. For instance, they might be unable to do so because of zoning regulations or restrictions of land availability. If this is indeed the case, then the model derived above will provide multiplier estimates that are unrealistically large due to expectations regarding supply response. A more reasonable assumption is that the availability of land restricts the production of goods and services. Therefore, the standard I-O model needs to be modified to incorporate supply constraints on certain production activities, permitting a more realistic evaluation of multiplier effects of injections into the economy. To account for restrictions in supply, several authors have developed models with supply assumed to be completely inelastic in some of the sectors (Lewis and Thorbecke, 1992; Miller and Blair, 1985; Parikh and Thorbecke, 1996; Subramanian and Sadoulet, 1990).

$$\begin{bmatrix} X_{no} \\ Y_{co} \end{bmatrix} = \begin{bmatrix} P0 \\ R - I \end{bmatrix}^{1} \begin{bmatrix} \overline{Y}_{no} \\ \overline{X}_{co} \end{bmatrix} \qquad (12.3)$$

where the sub-matrices are as follows:

P the $k \times k$ matrix containing the elements from the first k rows and the first k columns in $(I - A)$; the sectors have been labeled so that the first k sectors indicate the endogenous elements and the last $(n-k)$ sectors are the exogenous sectors. P is a matrix representing average expenditure propensities of non-supply constrained sectors.

R the $(n - k) \times k$ matrix containing elements from the last $(n - k)$ rows and the first k columns of $(I - A)$; R is a matrix representing average expenditure propensities of non-supply constrained sectors on supply constrained sector output.

X_{no} the k-element column vector with elements x_1 through $x_{k;}$, representing endogenous total output of non-supply constraint sectors.

Y_{co} the $(n - k)$-element column vector with elements y_{k+1} through y_n, representing endogenous final demand of supply constraint sectors.

Q the $k \times (n-k)$ matrix of elements from the last $(n-k)$ rows and first k columns of $-(I - A)$; the matrix Q represents supply constrained sector expenditure propensities on non-supply constrained sector output.

S the $(n - k) \times (n - k)$ matrix of elements from the last $(n - k)$ rows and columns of $-(I - A)$; the S represents here a matrix of average expenditure propensities among supply constrained sectors.

$\overline{Y}_{no}$ the k-element column vector of elements y_1 through y_k, representing exogenous final demand for non-supply constrained sectors.

$\overline{X}_{co}$ the $(n - k)$-element column vector of elements x_{k+1} through x_n, representing exogenous total output for supply constrained sectors.

In the modified model, changes in exogenous final demand for non-supply constraint sectors or changes in exogenous supply for the constraint sectors are met by changes in output for the unconstraint sectors and by changes in imports and exports for the constraint sectors.

The derived potential net export of the products of supply-constraint sectors (T) is the difference between the exogenous and endogenous final demand in the corresponding sectors:

$$T = \overline{Y}_{co} - Y_{co} \, . \qquad (12.4a)$$

Exogenously generated potential output (x_f) is calculated by dividing the land per land-use category available in 2025 ($\overline{L}_f$), which includes agricultural land, grassland, and forestland, by the respective future land requirement coefficient (c_f):

$$x_f = \overline{L}_f / c_f \; . \tag{12.4b}$$

The AEZ assessment model is used to derive regional differences for the land requirement and land productivity coefficients, for the disaggregation of the agricultural sectors into six sub-sectors in each regional I-O model, and for the calculation of exogenously generated potential output.

The AEZ method was developed by IIASA and the Food and Agriculture Organization of the UN (FAO) (FAO, 1995). It was repeatedly used and subsequently improved in several global and national studies (FAO/IIASA, 1993; Fischer and Makowski, 2000). The AEZ algorithm assesses the potential suitability and productivity of a particular land area for agricultural uses, depending on its soil, terrain, and climate conditions and at given input and management levels (see also the contribution of Fischer and Wiberg, Chapter 16).

The strength of the AEZ method is manifested in its ability to match land quality with the ecological requirements of the respective plants for soils, climates, etc., under explicit recognition of the socio-economic setting. The application of this method allows us to calculate this part of regional differences that is basically determined by natural factors. We apply the results of the AEZ assessment for the sectors of grains, other crops, and pasture livestock production. Due to the fact that land suitability changes along with changes of different land utilization types prescribed by certain social and economic conditions, three production scenarios for low, medium, and high input levels are developed (see, e.g., Xie and Jia, 1994). Variations in input levels are represented by the differences in multi-cropping indices; scale and intensity of land management; factor-intensity of labor, capital, and energy utilization; and operational technologies employed.

12.3 China and the Regions: Representation of the Economy and its Land Base

12.3.1 China's economy

In our I-O model, China is divided into 8 regions on the basis of their unique geographic, agro-climatic, demographic, and economic development levels. They are consolidated with provincial-level administrative boundaries for the sake of data availability and consistency. These eight regions are presented in *Figure 12.1*.

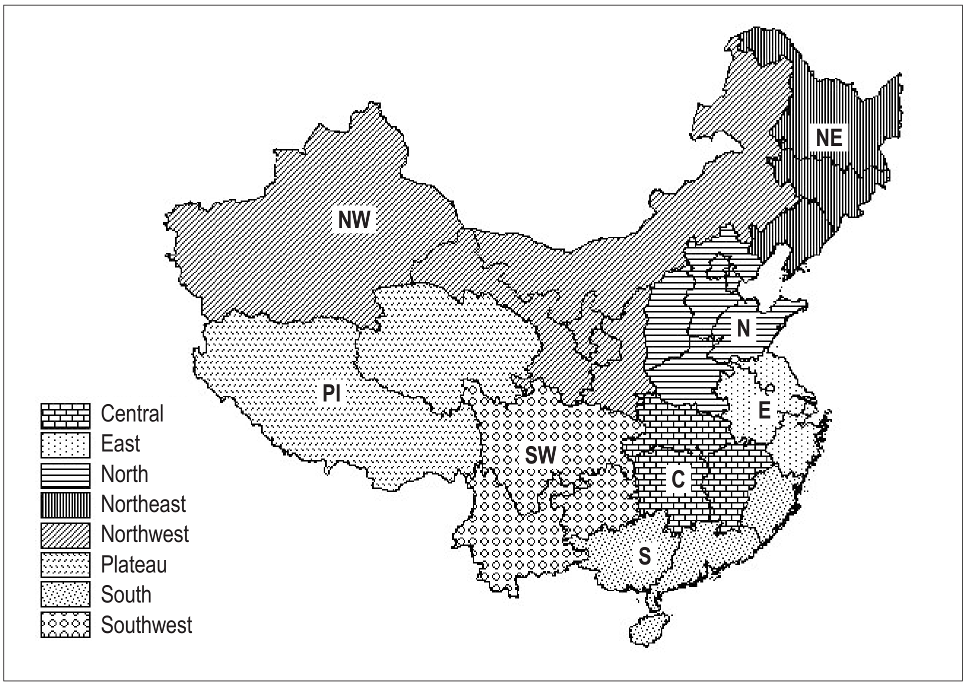


Figure 12.1. Map of China with provincial boundaries and the eight economic regions.

Seven out of the eight regions form the geographic building blocks in the I-O model. The Plateau region is not considered because of a lack of I-O data. Although it holds strategic importance in terms of geography and politics, its economic shares in the national economy are minor in comparison to the main I-O indicators of the other regions.

The economy of China and its regions is represented by the 1992 I-O tables. These existing tables were constructed by the Department of National Economic Accounting within the State Statistical Bureau of China (SSB, 1996 and 1997). The national table includes 118 sectors, with 6 agricultural sectors; however, the regional table exists only in a more aggregate form with only one agricultural sector. In order to obtain the desired I-O tables for each region, we first apply a procedure based on adjustment of national coefficients using the techniques of *location quotients* (LQs) adjustments. Then, we minimize the sum of squares of the percentage difference between the unknown cell figures and those obtained from the LQs procedure for each regional table, subject to the typical I-O balancing condition and other given sum requirements (for more detail see Sun, 2000).

12.3.2 Projections of future technology

The impact of changes in the economy and society on land use will depend on patterns of consumption as well as production. Extent and patterns of consumption are discussed in the form of various scenarios in Section 12.4. The patterns of production are represented in the technology matrix or A-matrix. Their immediate effects on land use are represented in the land-requirement coefficients or C-matrix [see Equation (12.2a)]. In order to project the future production functions of the respective sectors and the related effect on land use, we use a mixed approach of applying case studies and the RAS method[1] (see *Table 12.1*). We use the case studies for projecting key cells of the future production functions of certain sectors and calculate the missing cells with the RAS method (see, for example, Budavari, 1982, p. 404; Miller and Blair, 1985).

The case-study methodology was suggested and applied by Duchin and Lange (Duchin *et al*., 1993; Duchin and Lange, 1992 and 1994). The purpose of this approach is to develop a number of scenarios about the future regarding certain key economic sectors in terms of growth and technologies and to construct a corresponding database that contains the quantification of these parameters (Idenburg, 1993). The development of such case studies requires assembling information from many sources, such as technical publications and databases, and expert opinions. Time and budget constraints prevented us from conducting such case studies with technical details. As a sound compromise, we selected those variables for the case studies provided we had relatively reliable data for them. These additional data involved partial or full information on land inputs, the intermediate purchases and deliveries, value added, final demand, and the total outputs. The remaining missing data were estimated by the RAS procedure.

12.3.3 Land use in China

The I-O model is extended to incorporate land use. The land-use data are derived from the IIASA–LUC database. A number of fairly large and detailed geographical databases on China, including biophysical attributes of land and statistical data at the county level, have been incorporated into the LUC geographical information system. These data sets allow us to estimate the land acreage used in each of the economic sectors.

[1]The term RAS refers to a mathematical procedure for adjusting, sequentially, rows and columns of a given I-O coefficient matrix, $A(0)$, in order to generate an estimate of a more recent matrix, $A(1)$, when only the new structural information of sectoral output, $X(1)$, intermediate deliveries, $U(1)$, and intermediate purchases, $V(1)$, are assumed known. Once the procedure converges, the final outcome used to be denoted as $A(1) = RA(0)S$, in which R is a diagonal matrix that is the product of a series of diagonal matrices, and so is S.

Table 12.1. Scheme of I-O table of China in 2025.

	Grains	Other crops	Forestry	Live-stock	Handi-craft	Fishery	Industry	Construc-tion	Trans-port	Trade	Services	Int. del-liveries	Final demand	Total output
1. Grains				B			B					U	FD	X
2. Other crops												U	FD	X
3. Forestry												U	FD	X
4. Livestock												U	FD	X
5. Handicraft												U	FD	X
6. Fishery												U	FD	X
7. Industry				B								U	FD	X
8. Construction												U	FD	X
9. Transport												U	FD	X
10. Trade												U	FD	X
11. Services												U	FD	X
Intermediate purchases	V	V	V	V	V	V	V	V	V	V	V			
Value added							X–V							
Total output	X	X	X	X	X	X	X	X	X	X	X			
Land in Yuan/ha	L	L	L	L	L	L	L	L	L	L	L			

Notes: Ls are derived from literature and the AEZ model. Us, Vs, and Xs are derived for the major economic sectoral groups of agriculture, industry, and services from World Bank estimates and by comparison to structural changes in industrialized countries over a longer time period. Sub-sectoral shares within the agricultural sector are derived from an AEZ-based scenario assessment. Bs are subject to a restricted lower-bound in the optimization procedure, respectively, so as to guarantee a sufficiently high figure in the corresponding cell, which would partly reflect the increasing share of feeding mode in livestock production.

Table 12.2. Land requirement coefficients for China's regions in 1992 and 2025 (ha per million Yuan).

Economic regions	Grains		Other crops		Forestry		Livestock	
	1992	2025	1992	2025	1992	2025	1992	2025
R1–North	363.8	249.5	145.0	102.6	510.7	382.99	330.2	104.7
R2–Northeast	513.3	366.5	169.5	75.0	8,663.3	6,497.48	1,132.4	353.2
R3–East	236.8	183.1	103.1	83.5	730.7	548.00	39.9	12.6
R4–Central	231.9	178.3	105.1	64.9	1,438.8	1,079.10	434.7	136.2
R5–South	326.6	226.7	76.4	50.6	1,112.0	833.97	443.9	141.8
R6–Southwest	450.1	329.5	149.8	110.0	2,599.3	1,949.45	2,303.0	752.0
R7–Northwest	786.5	517.5	233.7	154.7	5,387.5	4,040.62	24,608.5	17,774.6
China	391.3	281.8	130.9	101.0	2,088.9	1,566.69	2,928.0	2,661.8

Sources: Figures for 1992 are calculated based on regional I-O tables. Procurement prices for grains and other crops are taken from *Price Statistical Yearbook of China* (1992, pp. 302–334). Assessment of crop production potential is taken from Xie (1994). Regional variation of livestock production is based on Zheng and Tang (1994).
Notes: We assume that land losses due to erosion could be fully compensated by reclamation. Land losses due to development of other economic sectors or residential use are subsumed in the category built-up land and subtracted from the other categories.

These data, together with that provided by the I-O tables, allow us to calculate land requirement coefficients (thus land productivity coefficients), for the base year. The current land requirement coefficients shown in *Table 12.2* represent average productivity of the total acreage in a given land use category. The use of these coefficients in scenario analysis would give us the land requirement at present-day efficiency. The higher the number in each cell, the less productive is the land to produce the respective output. The huge variability of coefficients for livestock production is partly due to the varying shares of pasture versus farm-based livestock production across regions, and partly due to the different environmental factors such as soil, temperature, and precipitation, which influence grassland productivity significantly.

12.3.4 Projection of land availability in 2025

Land availability forms a binding constraint to land-use requirements in general and for agricultural land uses in particular. Without additional available land, the only choice left for an economy is either to increase land productivity or increase imports. Given the foreseen scenarios of land productivity improvement and land availability, the balancing of the I-O model will give the required net import of land-based sectors so as to meet the additional final demand created by changes in the economy and society. Stated differently, given land availability, there is a clear-cut trade-off between land productivity improvement and net import requirement in scenario designs.

Due to sharp increase in land scarcity in recent years, we may expect that more efforts will be made to increase land reclamation and to protect agricultural land. As a consequence, we assume that the degradation-induced total losses of cultivated land, grassland, and forestland between 1992 and 2025 could be fully compensated by land reclamation and preservation. This assumption reflects also the policy orientation of the Chinese government. Nevertheless, the land conversion from agricultural uses to higher value-added nonagricultural uses and to residential uses will certainly continue. This conversion will take place mainly around economic centers. To capture this conversion, we employ the GIS technique to calibrate our scenarios. We overlap the map showing existing agglomeration with the map containing current land uses. We expand existing agglomerations by adding an additional ring of a certain width to the outskirts of each existing built-up area. The determination of this width is based on the scenarios of future demand for residential and nonagricultural uses of lands. In this way,[2] we can see how the expansion of existing built-up areas reduces the amount of other land-use categories.

For the calculation of land requirements per land use category, we need a land distribution matrix [R in Equation (12.3)] as shown in *Table 12.3*. This matrix establishes the linkage between land uses by economic sectors and natural categories of land. The entries in *Table 12.3* are numbers between 0 and 1, which indicate the percentage distribution of land used, by economic sector, in each of the major land categories. The numbers do not represent current patterns of land use but rather future land-use development. As the table shows, we assume that various land-use options, such as residential land, industrial land, horticulture, and fish, compete for cultivated land, grassland, and forestland. The category of unused or multiple-use land represents a residual value. In the case of fish production, for example, part of it is farmed on agricultural land without diminishing the usage of agriculture land. This sort of multiple use does not decrease the ability to use land for other production purposes. Sectors using built-up land are assumed to expand also, in part, on previously unused land. The land requirement coefficients based on the available land are shown in *Table 12.2*.

12.4 The Driving Forces of Land-Use Change

During the post-reform economic growth period after 1978, China has experienced a dramatic loss of arable land. Research has shown that both industrialization and

[2]The existing urban areas are captured with remote sensing. Unfortunately, this method only recognizes built-up areas beyond a certain size. As a consequence, the so-derived land conversions reflect only the extension of larger agglomerations.

Table 12.3. Regional distribution matrix in 2025.

Economic Sectors	Major land categories: Grains	Other Crops	Forest-land	Grass-land	Water areas	Unused	Total
Grains	1	0	0	0	0	0	1.00
Other crops	1	0	0	0	0	0	1.00
Forestry	0	0	1	0	0	0	1.00
Livestock	0	0	0	1	0	0	1.00
Fishery	0.10	0	0	0	0.89	0.01	1.00
Developed[a]	0.61	0.11	0.13	0.07	0.07	0.08	1.00

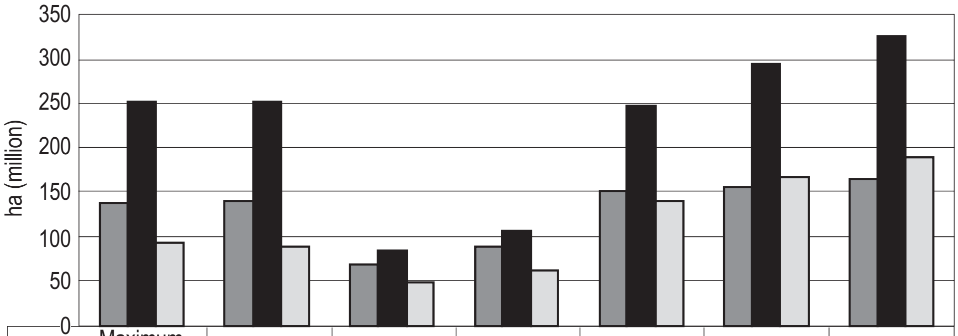
[a] The "Developed" category includes residential land, infrastructures, and industrial and commercial uses.

Source: Based on scenarios in the GIS: expansion of existing agglomerations by adding an additional ring of one kilometer width, to the outskirts of each existing built-up area.

land degradation have played an equally important role in reducing the total acreage of arable land (Sun and Li, 1997). Dynamic changes in economy and society are expected to effect further losses of cropland in coming decades. For a comprehensive understanding of the future land-use dynamics, we have identified a number of major driving forces represented by six scenarios (see lower part of *Figure 12.2*).

After establishing scenarios for each of the major driving forces (technological change, population growth, income growth, and urbanization), we introduce step by step the additional effects caused by each scenario. Starting from the base year reality, a set of scenarios representing each of the major forces is added to show its additional effects on land requirements. Scenario A represents the real situation in the base year 1992, with the technology and population level, share of urban and rural population, consumption pattern, and economic structure of 1992. Scenario B applies plausible technology available in year 2025 to the socioeconomic and demographic structure of 1992. In Scenario C, we add to Scenario B final demand changes and additional direct land requirements caused by a population of 1.49 billion people. Scenario D includes per capita income growth as well as lifestyle changes as represented by a set of income elasticities. Scenario E deals with the aggregate effects of Scenario D plus urbanization. Scenario F is designed to see the overall effects of a higher population estimate of 1.55 billion people and a higher share of urbanization combined with Scenario E.

The sections that follow discuss each of the driving forces affecting land use: economic and income growth, population growth, urbanization, changing consumption patterns, and technical change and land productivity.



	Maximum available land in 2025	A	B	C	D	E	F
Cultivated land	137	141	67	88	151	156	166
Grassland	252	252	85	108	247	293	326
Forestland	94	88	49	63	140	168	188

Major driving forces	Scenarios A (China 1992)	B (A+	C (B+	D (C+ income	E (D+	F (E+
Technological change	1992	2025	2025	2025	2025	2025
Population (in billions)	1.171	1.171	1.49	1.49	1.49	1.55
Income growth	1992	1992	1992	2025	2025	2025
Urbanization	1992	1992	1992	1992	2025	2025[a]

[a] Higher urbanization rate of 59% for 2025 (Shen and Spence, 1996).

Main assumptions: B: A+ Annual land productivity gains of 1%, 1.38%, and 0.68% for cropping, livestock, and forestry, respectively. C: B+ population of 1.49 billion. D: C+ 4.2 to 5.7% average annual growth rate of per capita income with the associated income elasticities (thus lifestyle). E: D+ 50% urban population with the associated expenditure patterns. F: E with population of 1.55 billion. Urban and rural infrastructure, residential land, and services are linked to a set of land per capita ratios, industrial land is assumed to stay constant. In all of the scenarios, trade balances of land-intensive products are kept proportional to today's imports and exports.

Figure 12.2. Land requirements of different scenarios.

12.4.1 Economic growth and the consequent per capita income growth

Since 1978, China's GDP has expanded at an average rate of nearly 10% – and total exports at 17% – per year. The Fifteen-Year Perspective Plan (1995–2010) identifies two fundamental transitions to sustain future growth: 1) movement from a traditional planned economy to a socialist market economy; and 2) transition from the extensive growth path, based on increases in inputs, to an intensive growth mode, driven by improvements in efficiency. Measures to sustain further growth include the restructuring of large state-owned enterprises; promoting science and

Table 12.4. Annual income growth rates in China for the period 1992–2025.

Regions	1992–2004		2005–2025	
	Rural	Urban	Rural	Urban
East Region: R1–R5	0.0475	0.0525	0.0425	0.0475
West Region: R6 and R7	0.0450	0.0500	0.0400	0.0450
China	0.0470	0.0520	0.0420	0.0470

technology; developing machinery, electronics, petrochemicals, automobiles, and construction as the pillar industries; and stimulating the growth of basic agricultural products, especially grain, cotton, and oilseed (World Bank, 1997b). Assuming the continuance of high saving rates supporting high investment rates, market-oriented reforms, and high factor productivity growth, the World Bank projected growth rates of 6.6% annually until 2020. The projections for individual sectors range from 3.8% for agricultural sectors, to 6.6% for industrial sectors, to 7.6% for service sectors (World Bank, 1997b, p. 21). According to the World Bank, the pace of GDP growth will slow down over time, from some 8% today to 5% in 2020, due to a, by then, stagnating labor force, diminishing marginal returns, and lower gains from structural change.

These aggregate growth trends mask diverging paths for different parts of China. There is a large body of literature dealing with the regional disparity in China (Liu *et al.*, 1999; Tian, 1999, among others). It is generally acknowledged that three regions have emerged with discernable development paths in the past two or more decades: 1) the leading coastal areas, characterized by high income level and high growth rate; 2) the central regions, with an average income level, but catching up with rapid structural changes from agriculture to industry and services; and 3) the backward regions in the west, with a much slower growth rate, and with a small share of the population dominated by national minorities. Another significant disparity exists between rural and urban areas. The per capita income ratio of rural to urban residents has been around 1 to 2.5 in the past two decades.

GDP growth rate is a comprehensive indicator that is not independent of population growth (implying labor force growth) and technological progress. To make income growth rate be independent of other driving forces, we subtract the foreseen growth rate of population and the part corresponding to technological progress (about 35% of GDP growth) from the predicted national GDP growth rate (World Bank, 1992). As a result, we obtained a net per capita income growth rate, and we call it that, for simplicity. In order to accommodate to the regional and rural versus urban differences discussed above, we distinguish growth rates for urban and rural areas and for two large development zones: the East Zone composed of Regions 1–5 and the West consisting of Regions 6 and 7. Finally, we have the basic scenarios for per capita income growth as presented in *Table 12.4*.

12.4.2 Population growth

When the People's Republic of China was founded in 1949, it had a population of 540 million; three decades later its population was more than 800 million; and present China's population has approached 1.3 billion. Today's high share of young Chinese at reproductive age has created a strong population momentum that is now driving China's population growth despite already low levels of fertility. China is confronted with two counteracting trends: while economic growth, urbanization, and the associated lifestyle change may lead to lower fertility rates, modernization and the opening of society might lead to opposition to the government's strict one-child rule in family planning (Heilig, 1999). In its most recent (medium variant) projection, the UN Population Division estimates that China's population will increase to 1.49 billion in 2025 and then slightly decline to 1.488 billion in 2050 (United Nations Population Division, 1998). A somewhat higher projection estimates 1.55 billion people for 2025 (Shen and Spence, 1996).

A crucial characteristic of China's demographic situation is the concentration of its large population in the eastern part of the country, especially in the coastal zone. A large part of China's land is virtually uninhabited, such as the Gobi Desert, the steep slopes of the Himalayas, and the vast dry grasslands of the northcentral region. Roughly 1.1 billion people (or about 90% of the population) live in only a little more than 30% of China's land area. The population density of this area is 354 people per square kilometer (km^2). The skewed spatial distribution of the population is a consequence of the country's uneven distribution of agro-climatic and bio-physical environments, as well as the uneven pace of industrialization.

In the past two decades, two opposite trends have coexisted to shape the population dynamics across regions. On one hand, migration from Western and Central China to the eastern regions, especially the coastal areas, adds percentage points to population shares of the eastern regions. However, on the other hand, the fertility rates moving upward from the eastern to the western regions have basically counterbalanced, if not exceeded, the impact of migration (Jiang and Zhang, 1998). In addition, one must consider the moving of traditional industries – particularly heavy industry – from the eastern regions inward toward the western regions and the new strategic movement of the Chinese government to reduce regional disparity. Due to the counteracting tendencies of the various trends in the coming decade, their accumulative impact up to 2025 may not be very significant. We assume that the population shares of the East and South Regions, the most developed regions, in the national total will increase by one percentage point; the population shares of Central and Southwest Regions, the regions with high population density and the highest proportion of agricultural population, will decrease by one percentage point; and the population shares of other regions will stay unchanged.

12.4.3 Urbanization

Despite the fact that the urban population is constantly increasing, China can still be considered a predominantly rural society. In 1997, after the rapid increase of the officially defined urban population for more than a decade, only some 30% of the population lived in urban areas. The rather recent increase in urban population is mainly due to the promotion of towns into cities, an action that increases the number of cities altogether. The number of cities is growing rapidly. In 1980, there were 223 cities throughout China; by 1990, the number had more than doubled to 467. In the last 10 years, the number of large cities has increased from 70 to 119, small cities from 108 to 289 and towns from 2,874 to 12,084 (Heilig, 1999). Another reason for the increase in urban population has been the loosening of strictly controlled internal migration to meet the labor demand of the growing cities and towns. In addition, recent years have seen a wave of temporary "illegal" rural-urban labor migration, called the "floating population." Some estimates put the size of the floating population in large cities at as much as 25% of the urban population (Heilig, 1999). We assume that this urbanization trend will continue, and that by 2025 about 50% of the Chinese population will live in urban areas. This assumption is consistent with the corresponding UN projection (United Nations Population Division, 1998). We further assume that this agricultural population living in cities and towns will gradually adopt urban lifestyles.

There are no reliable estimates of the urbanization rate for different regions, since even present data on city growth and rural-urban migration is of poor quality. However, as we discussed in Section 12.4.2, two large zones can be distinguished due to the striking development disparity between them. For the more developed eastern zone (Regions R1–R5), we assume urbanization rates of roughly 54%, and for the less developed western zone (Regions R6 and R7), about 44%, respectively.

12.4.4 Change in consumption patterns

With respect to changes in consumption patterns, changes in diet structure are the most relevant to the study of land-use change. In China's food tradition, cereal products have been of overriding importance. Other food products such as meat, fishery products, vegetables, and fruit played only a residual role in human diet. This pattern has been changing due to recent social and economic developments. Urban residents typically prefer a more diverse diet and eat more processed foods. Today's Chinese eat more meat and dairy products, which has boosted livestock production. China's population has enormously increased its meat consumption and also eats more fruits and vegetables, whereas direct consumption of grain has leveled off or even declined. For example, over the period from 1981 to 1995, the direct food grain consumption per capita in urban areas[3] dropped from 145 kilograms (kg) to about 100 kg, whereas in rural areas, the per capita consumption of milled grain increased first in the early 1980s and gradually went back to the 1981 level by 1995 (Wu and Findlay, 1997, p. 49). Despite these developments, China's average food calorie supply per person per day[4] is still below the average level of developed countries (FAOSTAT, 1998). Therefore, an increase in per capita calorie consumption can be expected in the future.

A comparison of per capita calorie intake across some representative countries shows that today's food calorie supply of animal products in China is about 467 kilocalories (kcal) per person compared to 503 kcal in South Korea, 600 kcal in Japan, and 1,006 kcal in the USA. The average consumption for developed countries is 867 kcal daily. In addition, today's calorie intake of fish in China is behind other Asian countries. Currently, food calorie supply of fish in China is 29 kcal, compared to 92 kcal in South Korea and 194 kcal in Japan (FAOSTAT, 1998).

To incorporate these considerations in a consistent way and in line with our I-O modeling, we established the scenarios of income elasticities for two periods, 1992–2005 and 2005–2025. We based them on the estimates and calibrations in Huang and Rozelle (1998) and Huang and Chen (1999), making our own adjustments. Combining the scenarios of per capita income growth with those of income elasticities gives the scenarios of per capita expenditure pattern for the year 2025.

To calculate aggregate final demand from households for the products of each production sector, we multiply the average expenditures of urban and rural residents, respectively, by the total numbers of urban or rural residents in each region. To obtain total final demand corresponding to each production sector, we link other final demand components to household consumption according to their current ratio to the level of aggregate household consumption.

12.4.5 Technical change and land productivity

China's economic success has been accompanied by enormous productivity gains. Yet China's economy is still characterized by substantial inefficiencies and backward technologies and therefore has enormous possibilities for improvements.

[3]There is an increasing extent of under-reporting of food consumption for urban households in the reform era, because eating in restaurants and at the workplace has become increasingly popular and fashionable. The official household survey has a limited ability to incorporate this trend fully. If this trend is taken into consideration, the decrease of per capita grain consumption in urban areas may not be so significant.

[4]Estimates of future demand for meat are difficult to make. The vast differences in the estimated results are directly related to the different parameters and research methods adopted in different studies. Furthermore, great inconsistencies of the data on meat consumption and output exist due to a combination of reported data on the supply side and survey data on the demand side (Feng, 1997).

The World Bank (1997b, p. 20) estimated that annual growth rates of total factor productivity of 5% to 7% during 1995–2020 will lead to major changes in the relative size of economic sectors in terms of both output and labor force. For example, agricultural employment is expected to fall from more than half of total employment today to one-quarter within the next 25 years.

These structural changes in the economy will have considerable effects on the structure of land use. Most important for our question of land-use change in China are technical changes and productivity gains in the primary sectors. In grain production, average yields in China are generally higher than in developing countries, but still well below the averages in developed countries.[5] In the future, farmers could boost grain production via a significant growth in yield, by planting more updated hybrid seeds; balancing the use of chemical fertilizer and pesticides; increasing the use of other modern inputs such as plastic film, farming machines, and power for agriculture use; investing in agricultural infrastructure such as irrigation and drainage facilities; and conducting agricultural research (Lin, 1995; Lin *et al*., 1996; Nickum, 1982; World Bank, 1985 and 1997a).

In addition, the Ministry of Agriculture plans to classify over 80% of farmland as basic farmland conservation zones by 2010. The average increase in land productivity in the grain production for the period of 1950–1977 was 3.1%. There is a debate on the magnitude of the future performance of grain production.

Estimations of annual yield increases vary between 0.5% and 2% depending on assumptions made about investment in research and irrigation, world price impact, salinity and erosion, and opportunity costs of labor and land (Cao *et al*., 1995; Huang and Kalirajan, 1997; Lin, 1995; Lin *et al*., 1996; World Bank, 1997b).

For our scenarios, we follow the Agricultural Action Plan of China's Ministry of Agriculture with an increase in the target of grain yield per unit area of 1% per year. We apply the same productivity growth rate for other crops.

In order to derive regional differences, we use an assessment of the crop production potential in China by Xie and Jia (1994) based on the AEZ method. This approach allows us to calculate regional differences based on natural factors, assuming similar technologies in all of China. Because land suitability changes along with changes in different land-use types prescribed by certain social and economic conditions, Xie and Jia (1994) developed three production scenarios at low, medium, and high input levels. The differences among these three input levels can be attributed to differences in multi-cropping indices; scale and intensity of land; pest and weed management; factor-intensity of labor, capital, and modern energy use; utilization of organic and chemical fertilizers; and other operational technologies employed.

[5]It should be noted that the current consensus in China is to promote high-value-added and labor-intensive agriculture rather than the Western type of capital-intensive agriculture.

The productivity of grasslands in China is much lower than in other parts of the world (Chen and Fischer, 1998), which severely limits the development of China's livestock industry. Officials (Ministry of Agriculture, 1999) state that China has a serious problem of grassland degradation: more than 50% of the northern grasslands are degraded and the rest is degrading at the rate of 1.9% annually. To maintain and improve the grassland quality, the Ministry of Agriculture plans to apply measures such as pest and rodent control, monitoring, conservation zones, and enclosed pastures. The improvement of 25 million ha of pasture is planned by 2010. The Ministry of Agriculture (1999) hopes that China can maintain a stable output of animal husbandry in pastoral areas before 2010 and can start to increase productivity of pastoral land afterward. Given the very limited capability of the grazing mode, it is widely expected that by 2020, a feeding mode would produce more than 80% of the total livestock output (Ministry of Agriculture, 1999). We put this figure at about 84% for 2025. In line with the expectation of the Ministry of Agriculture (1999), we assume an accumulative land productivity growth of 25% for the whole period of 1992–2025 in the pasture sector.

In calculating land requirement coefficients for the livestock sector, we exclude the indirect land uses for growing processed and unprocessed feed-crops, to avoid double counting. However, we include those land uses for keeping pork and poultry, which are not grassland and amount to a small share of residential land. Because it is meaningless to talk about land productivity growth of the feeding mode based on those land uses for keeping animals, we use an output-weighted combination of pasture land productivity growth and feed-crop land productivity growth to define the overall land productivity growth in the livestock sector. Because the major feed crops in China are maize and soybeans and their yield growth has been higher than that of all grains, we assume an average annual growth rate of 1.25% for the productivity of feed-crop land.

To derive the regional variation on land productivity growth in the livestock sector, we use the estimate by Zheng and Tang (1994). Their assessment of grassland productivity consists of two major parts: the calculation of the primary production (forage output) and the secondary production (livestock products). The calculation of the forage is based on the AEZ method. To receive the feed supply potential, the yields for each zone were adjusted by feed intake rates of livestock and different dehydration rates for hay – representing different grassland types (Zheng and Tang, 1994, p. 33). In a second step, the characteristics of the livestock system were highlighted and different herd proportions and livestock production potentials were calculated based on the balance between feed supply potential and feed requirements.

The incremental output, from the low-input assessment to the intermediate-input assessment, shows a clear variation across regions. We use this variation to distribute average national productivity-growth rate to the regional level.

In forestry, future development could be more seriously constrained. But the estimations of forest stock are quite diverse. For example, Fischer *et al*. (1996) report that the stock has steadily increased, from about 7 billion cubic meters (m^3) in the 1950s to about 10 billion m^3 in the late 1980s. Liu (1998) claims that timber stocks are drastically decreasing due to increased consumption, withering, fire damage, and insect damage. For 1992, Liu estimates the existing forest stock at about 5.3 billion m^3 and states that if no action were taken, China would lose all its timber stocks in the near future. Nilsson (1999) shows that the felling of industrial wood at the current rate of 197 million m^3 per year exceeds the annual increment of 176 million m^3 per year in growth of natural forests and industrial plantations. Shi and Xu (2000), on the other hand, using data from all four forest resource censuses (Ministry of Forestry, 1978; 1983; 1990; 1996), show that the forest resource stock had been slightly increasing between 1973 and 1993, and that the total timber stock was about 11.8 billion m^3 in the early 1990s.

On the basis of the forest resource census data, which are relatively reliable and authoritative, we establish the productivity change scenarios for the forestry sector in line with Fischer *et al*. (1996) and Shi and Xu (2000). We first calculate the land requirement coefficient in the forestry sector for 1992 directly from the total output of forest sector as provided by the I-O tables and the corresponding areas. The resulting land coefficient at the national level is consistent with a sustainable yield factor of about 4.2 m^3/ha (Ministry of Agriculture, 1998; Shi and Xu, 2000), while being weighted by the corresponding major price figures of forest products.

Since timber densities in China are very low (30 m^3/ha to 84 m^3/ha in comparison to the world's timber densities of about 100 m^3/ha) (Ministry of Forestry, 1990) and efforts to improve forest management through property right reform, strengthening monitoring and preservation institutions, and employing other effective management practices are under way, we assume accumulated productivity growth of 25% in this sector from 1992 to 2025. Due to the lack of AEZ assessment, we have to ignore the regional variation in growth for this sector.

Estimation of future land areas required for urban development, rural industrial agglomerations, and infrastructures is less straightforward. The heterogeneity of industrial production allows no systematic aggregate data on land requirements of various industries beyond case studies on a local or regional level (e.g., Borchard, 1999). In addition, international data usually includes commercial and industrial land with various shares of services and industrial production. Much of the infrastructure is already in place, so that the improvement and extension of this infrastructure would only require marginal additions of land. Future development might mainly necessitate a restructuring of existing areas and infrastructures. It is difficult to say how redevelopment of urban areas and organizational changes will affect land productivity in the industrial sectors. Given the backward structure of Chinese industry, there might be a great possibility to have considerable increases in industrial value-added without any significant additional industrial land use, simply because most of this growth will be outside the traditional smoke-stack, heavy industry (and land-use intensive) activities.

Even if one assumed that the industrial areas would stay more or less the same, considerable additional land will be required for infrastructure development. The shortage of infrastructure has been considered one of the main bottlenecks for future economic development (China's Agenda 21, unpublished; EAAU, 1997; World Bank, 1985). China's annual investment in the transport and other infrastructure sectors has been small in comparison to other countries. Major investments in the extension of the current structure as well as in increasing its efficiency are necessary. The ninth five-year plan targeted a 12% increase of roads, 17% of railways, 35% increase of waterways, and more than 100% increase for aviation capacity (Spear *et al*., 1997). There is no information available on longer-term infrastructure development plans. A good proxy for land use consuming infrastructure is the future increase of roads. Currently, China averages 1.1 kilometers (km) of roads per 100 km^2, in comparison to 7 km/100 km^2 in the USA and 4.7 km/100 km^2 in India, respectively. Even the better-developed coastal areas have only 2.5 km/100km^2. Only 23% are asphalt paved and most are in poor condition (CIA, 1999; EAAU, 1997, p. 228). Projections show that the number of cars might more than triple within the next 10 years (TEI, 1994; quoted in World Bank, 1997a). In our study, we distinguish the coastal regions from the inland and use two different growth rates of transportation infrastructure development for them in the period 1992 to 2025. For the coastal area, we assume a relatively high annual growth rate of 1.9%, which would bring coastal China to the road-infrastructure level of today's India. For Central China and Western China, we assume annual growth rates of 1.6%.

12.5 Model Results

Given the commonly excepted scenarios for demographic, social, and economic changes as well as technological progress, sectoral outputs would drive the associated land requirements to exceed the available land area. In other words, China would not be able to support the increased demand for land-intensive products with its land base without significant improvement in land productivity and/or increasing imports. The major results of our scenario analysis at the national level are presented in *Figure 12.2* (see Section 12.4, above).

Table 12.5. Necessary annual growth rates in land productivity (%).

Economic sectors	Assumed growth rate in "Technology 2025" (in %)	Necessary growth rate[a] (in %)
Grain	1.00	1.03
Other crops	1.00	2.04
All crops	1.00	1.28
Forestry	0.68	2.65
Livestock	0.98 (0.68)[b]	1.45

[a]These necessary annual growth rates will keep the demand specified in Scenario E within the land limits.
[b]Land productivity growth in the pasture sector.

In the category representing *cultivated land*, the scenarios that add, step by step, per capita income growth with the associated lifestyle change, urbanization, and higher population growth (D, E, F) to the previous scenarios are exceeding the limits of available land. The biggest jump in demand for farmland is triggered by the income growth scenario (D). The difference between Scenarios D and E indicates that given the income level, the land saving-effects of urbanization (more efficient use of infrastructure and residential land) would be offset by an increase in indirect demand for feed-grain triggered by the higher consumption of animal products.

In the case of *grassland*, the scenarios that add urbanization and higher population growth to the previous scenarios (E and F) are exceeding the available grassland areas. The biggest jump in demand for additional grassland is caused by per capita income growth with the associated lifestyle change (D), in particular, by the significant increase in per capita meat consumption.

Similar to the case of cultivated land, the demand for forestry products exceeding the available *forestland* appears in the most aggregate scenarios (D, E, F).

It is worth noting that the desired high growth rate for land productivity in the livestock sector reflects the strong desire for feed grain growth. This indicates that there is a trade-off between the productivity growth rate of the grain sector and that of the livestock sector.

Table 12.5 presents a comparison of commonly expected growth rates in the literature and those necessary to keep the demand, as specified in Scenario E, within the land limits. We can see from this table that the required land productivity growth rate for other crops, forestry, and livestock are considerably higher than the ones commonly assumed in the literature.

Given the constraint of the immobile land availability and the technological scenario "Technology 2025," we proceed at both regional and national levels to show how much net import will be required to meet the regional and national demand for

Table 12.6. Deficit or surplus of the major agricultural products at regional and national levels in 2025 (in million Yuan).

Economic sectors	R1–N	R2–NE	R3–E	R4–C	R5–S	R6–SW	R7–NW	7–Regions[a]	China[b]
Grains	14,744	20,331	11,456	19,522	10,402	12,154	11,565	100,173	18,106
Other crops	–8,457	8,538	–34,470	18,319	–22,431	–19,694	6,787	–51,408	–137,562
Livestock	–18,926	3,318	–41,014	16,046	–17,353	–2,719	3,151	–57,497	–104,133
Forestry	–5,163	751	–551	–19,174	–12,383	–85	–1,969	–38,574	–34,182
Sum	–17,803	32,938	–64,579	34,713	–41,765	–10,345	19,534	–47,306	–257,771

[a] Category "7-Regions" is the sum of Regions R1 to R7.
[b] Differences between the categories 7-Regions and China represent the combination of Plateau Region and the central government-run economic sector.
Notes: Minus "-" means deficit and positive numbers mean surplus.

grain, other crops, livestock products, and forest products. *Table 12.6* reports the derived net import requirements.

At the national level, the net import demand for livestock products is at a scale of more than 100 billion (1992) Yuan, being equivalent to the domestic production cost of 100 million tons of wheat. The net import demand for products of the other crop sector is at an even larger scale. Such a large-scale net import of animal and crop products would go beyond the limits of political feasibility and the capacity of the world food market. Therefore, these figures would indicate that if maintaining the moderate pace of technical progress across these three major land-use sectors, as assumed in Technology 2025, the other crop sector will compete with the grain sector for claiming much more land and the livestock sector will put much stronger pressure on the grain sector for feed-grain production than we have assumed before.

At the regional level, the highest requirements for net import of crop and livestock products correspond to the most developed regions East and South, closely followed by the better developed region North in the coastal zone. The traditional food export regions Northeast and Central continue to be the leading contributors to the national food pool, followed by the relatively backward region Northwest. The traditional food export region, Southwest, becomes one requiring a moderate net import of food. In terms of forest products, Northeast and the central government-run forest sector (including Plateau Region) continue to show their significant advantages in forest resource endowment, and the more industrialized regions in the coastal areas remain net import regions. Meanwhile, the Southwest region, which closely follows the Northeast region in terms of rich forest resources, ends up with a moderate demand for net import. For a recent assessment on the regional distribution of China's forest resources, see Albers *et al*. (1998).

Putting the seven regions together, their aggregate deficit in the livestock and other crop sectors can be balanced by their aggregate surplus in grain production.

This would mean that there was no agricultural surplus left to meet the demand of the central-government-run economic sector and to fill the food deficit of the Plateau Region, which has historically depended on food transfer from other regions to meet a large proportion of its food demand.

Both the above assessments of desired productivity growth and import strongly suggest that in order to meet the foreseen increase in demand for major agricultural products, China needs to make greater efforts to improve land productivity in the cropping, pasture, and forestry sectors. China also needs to modernize its feeding mode at a faster pace in order to increase both output and efficiency in the farm-based livestock sector.

12.6 Implications for Future Land-Use Change

In this report, a set of diverse scenarios has been developed to show how population growth, changes in lifestyles, levels of urbanization and migration, and per capita income growth during the next two or three decades might affect the demand for different types of land in China. Given the moderate pace of technological progress, as assumed in the scenario Technology 2025, the resultant increases in final demands and sectoral outputs would drive the associated land requirements to exceed the then available land area. All three land categories face shortages for the most-aggregate scenarios. If the traditional policy of grain and food self-sufficiency were maintained intact, to keep the farmland requirement feasible, an annual growth rate of land-productivity of about 1.28% would be required. Taking into account the implicit substitution between the productivity growth rate of the grain sector and that of the livestock sector, an even higher rate would be required to make a proper balance between these two sectors. On the other hand, it is widely believed that due to current inefficiencies and structural problems, land productivity in China's cropping agricultural sector may have ample room to significantly increase above current levels even without having to rely on future technologies but by further exploiting the best currently available practices (Ministry of Agriculture, 1998 and 1999). With the help of newly available technology in the next two decades, China's cropping agriculture would have a great potential to reach a higher land productivity growth than 1.3% per year.

To realize the desired productivity gains in the cropping sector, large investments are needed to develop additional and efficient water supplies by improving irrigation systems as well as by increasing efficiency of on-farm use of water. The most significant contribution of sustainable and extended irrigation will occur in the regions North and Northwest. According to a recent AEZ assessment over the water issue in China (Heilig *et al*., 2000), if geophysically feasible irrigation can be guaranteed in these two regions, irrigation would be able to boost the gross grain production potential by about 100 million tons in the North and by 45 million tons in the Northwest, as compared with purely rain-fed cropping.

In terms of increasing the application rate of fertilizer, new facilities using modern technologies for fertilizer production and imports will be necessary. Currently, China is importing some 20% to 25% of internationally traded fertilizer to meet about 30% of its fertilizer nutrient needs (World Bank, 1997a).

A well-known alternative to the policy of self-sufficiency is the import of grains. However, one bottleneck hampering grain import is the lack of infrastructure, such as port handling equipment and transit storage. Here, large investments are needed to improve grain-handling efficiency (World Bank, 1997a).

Further productivity growth is also required to compensate for loss and degradation of available cropland. Currently, much cropland converted to non-agricultural uses is used inefficiently by land-extensive development projects and "horizontal" expansion of urban agglomerates. Strict measures must be implemented and enforced to minimize construction-related losses of cultivated land. China needs concepts for infrastructure development that minimize land requirements, especially in the rapidly developing coastal provinces.

Given today's practice of livestock production, China will have great difficulties in meeting increasing demands for animal products. Transformation of livestock production has already begun, from the peasant's backyard feeding mode and sedentary grazing to medium and large-scale "factory" types of production systems. Dependence on modern feeding methods and genetic techniques introduced from abroad is growing. The shortage of grassland can be partly alleviated by increasing the already high share of feed-grain production, but doing so increases the pressure on cropland. The competition between livestock production and urban and industrial uses for grassland is less pronounced than in the case of cropland, because grassland is more remote from economic growth centers. However, the remoteness, prohibitive transportation costs, as well as lack of infrastructure pose severe hindrances for grassland to reach higher productivity rates.

The remoteness of the locations of major forests also makes it difficult to improve land productivity. In addition, excessive harvesting and insufficient regeneration have caused long-term deficits in stocking. The increasing needs for timber logs from a growing economy and the enormous demand for residential firewood will require increasing efficiency as well. Policies are needed to encourage such efforts as improving forest management, increasing forest densities, and maintaining stable political and institutional environments for long-term investment. While the importance of forestland for protecting cropland by fighting soil and wind erosion and desertification has already been widely recognized, active and massive responses are still missing. Direct investments in regional reforestation projects already show some visible and encouraging impacts at the national level (Albers *et al*., 1998, p. 27). The quality of these planted forests, however, is so poor that they will not play any significant role in timber production in the near future (Yin, 1994).

To summarize, we can state that there is ample room for further growth in land productivity in the major land-use sectors of China's economy. However, it is uncertain that sufficient investments for the agricultural sector will be provided. The political goal of self-sufficiency in (regional) grain production will certainly collide with its high opportunity costs, given the higher productivity of land and water for urban and industrial uses. Increasing shortages in land resources might lead to increasing interest in remote areas and to closer ties across the regions within China as well as to closer trading relationships with other parts of the world.

The results of our study must be viewed with caution. Even if future land requirements could be satisfied, we do not say anything about the sustainability of land use by the various economic sectors. Also, questions of diversity are not addressed. For example, decreasing diversity and elimination of old growth forests may be considered the most serious environmental problem in the forest sector (Rozelle *et al*., 1998). In the cropping agricultural sector, the reclamation and yield growth potential are largely dependent on irrigation and water control. At this point of the I-O analysis we do not make any estimation of water availability and its effects on future land-use changes. In addition, some productivity growth scenarios presented in this chapter imply a higher use of pesticides, fertilizer, fossil fuel, and equipment. In this regard, future work should also pay attention to the energy needs, the emission of greenhouse gases, and other pollutants arising from the various development and technical options of the primary sectors.

Another step for future work will be to incorporate the possible impacts of climate changes, as estimated by Fischer and Wiberg (Chapter 16 in this volume), into the I-O modeling framework. This means essentially a reassessment of the land-use coefficients and their variability across the various regions. There seems to be no technical obstacle to such an extension. The framework we present of an extended I-O model has proven flexible enough to address these additional questions.

References

Albers, H.J., Rozelle, S.D., and Guo, L., 1998, China's forests under economic reform: Timber supplies, environmental protection, and rural resource access, *Contemporary Economic Policy*, **XVI**:22–23.

Borchard, K., ed., 1999, *Feststellungen und Empfehlungen für eine zukunftsfähige Raum- und Siedlungsentwicklung*, Akademie für Raumforschung und Landesplanung, Hannover, Germany.

Brown, L., 1995, *Who Will Feed China? Wakeup Call for a Small Planet*, W.W. Norton, New York, NY, USA.

Budavari, P., 1982, Generalization of RAS method: Linear restrictions with strictly convex distance function (duality theory and algorithm), in *Proceedings of the Third Hungarian Conference on Input-Output Techniques*, Statisztikai Szemle, Vol. 60, Heviz, Hungary.

Cao, M., Ma, S., and Han, C., 1995, Potential productivity and human carrying capacity of an agro-ecosystem: An analysis of food production potential of China, *Agricultural Systems*, **47**:387–414.

Chen, X., Wang, Z., and Guo, J., 1996, *China's Grain Supply and Demand in the 21st Century*, Bulletin of the Chinese Academy of Science, Report No. 10, Beijing, China.

Chen, Y., and Fischer, G., 1998, *A New Digital Geo-Referenced Database of Grassland in China*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

China's Agenda 21, *White Paper on China's Population, Environment, and Development in the 21st Century*, The Administrative Center for China's Agenda 21, Beijing, China, unpublished.

CIA, 1999, *The World Factbook 1999*, Central Intelligence Agency, Office of Public Affairs, Washington, DC, USA.

Duchin, F., Hamilton, C., and Lange, G.-M., 1993, *Environment and Development in Indonesia: An Input-Output Analysis of Natural Resource Issues*, Associates in Rural Development and Institute for Economic Analysis, New York University for USAID, Jakarta, Indonesia.

Duchin, F., and Lange, G.-M., 1992, *Case Study Methodology*, Institute for Economic Analysis, New York University, New York, NY, USA.

Duchin, F., and Lange, G.-M., 1994, *The Future of the Environment: Ecological Economics and Technological Change*, Oxford University Press, Oxford, UK.

EAAU, 1997, *China Embraces the Market: Achievements, Constraints, and Opportunities*, Australian Department of Foreign Affairs and Trade, Barton, Australia.

FAO, 1995, *Livestock: A Driving Force for Food Security and Sustainable Development*, FAO, Feed Resources Group (R. Sansoucy), Rome, Italy.

FAO/IIASA, 1993, *Agro-Ecological Assessment for National Planning: The Example of Kenya*, Food and Agriculture Organization of the United Nations, Rome, Italy.

FAOSTAT, 1998, *Food Balance Sheets*, Food and Agricultural Organization, Rome, Italy.

Feng, L., 1997, Feed demand and its impact on the grain economy in China, in *Agricultural Policies in China*, Organisation for Economic Co-operation and Development, Paris, France.

Fischer, G., Liu, Y., Zhao, M., and Sun, H., 1996, *Land-Use Change in China: Land Resources and Contemporary Land Use*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Fischer, G., and Makowski, M., 2000, Land use planning, in A.P. Wierzbicki and M. Makowski, eds, *Model-Based Decision Support Methodology with Environmental Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Gardner, G., 1996, Shrinking fields: Cropland loss in a world of eight billion, *Worldwatch Paper*, Worldwatch Institute, Washington, DC, USA.

Heilig, G., 1999, *Can China Feed Itself?*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Heilig, G.K., Fischer, G., and Velthuizen, H. van, 2000, Can China feed itself? An analysis of China's food prospects with special reference to water resources, *International Journal of Sustainable Development and World Ecology*, **7**(3):153–172.

Huang, J.K., and Rozelle, S., 1997, Environmental stress and grain yields in China, *American Journal of Agricultural Economics* **77**(4):853–864.

Huang, J., and Chen, C., 1999, *Effects of Trade Liberalization on Agriculture in China: Commodity Aspects*, CGPRT Centre, Jalan Merdeka, Indonesia.

Huang, J., and Rozelle, S., 1998, *China's Grain Economy to the Twenty-First Century*, China Agriculture Press, Beijing, China.

Huang, Y., and Kalirajan, K.P., 1997, Potential of China's grain production: Evidence from the household data, *Agricultural Economics*, **17**:191–199.

Idenburg, A., 1993, *Gearing Production Models to Ecological Economic Analysis: A Case Study, within the Input-Output Framework of Fuels for Road Transport*, Faculteit der Besturrskunde, Universiteit Twente, Enschede, Netherlands.

Jiang, Z., and Zhang E., eds, 1998, *Population Forecasting at the Provincial Level for China*, China Population Publishing House, Beijing, China.

Lewis, B.D., and Thorbecke, E., 1992, District-level economic linkages in Kenya: Evidence based on a small regional social accounting matrix, *World Development*, **20**(6):881–897.

Lin, J.Y., 1995, Grain yield potential and prospect of grain output increase in China, *People's Daily*, March 10, Beijing, China.

Lin, J.Y., Huang, J., and Rozelle, S., 1996, China's food economy: Past performance and future trends, in *China in the 21st Century: Long-Term Global Implications*, Organization for Economic Co-operation and Development, Paris, France.

Liu, A., Yao, S., and Zhang, Z., 1999, Economic growth and structural changes in employment and investments in China, 1985–1994, *Economics of Planning*, **32**:171–190.

Liu, X., 1998, Adjusted forest accounts for China, *Ecological Economics*, **27**:283–298.

Miller, R.E., and Blair, P.D., 1985, *Input-Output Analysis: Foundations and Extensions*, Prentice-Hall, Englewood Cliffs, NJ, USA.

Ministry of Agriculture, 1998, *Chinese Reserve Resources for Agriculture*, Regional Planning Office for National Agricultural Resources, Beijing, China.

Ministry of Agriculture, 1999, *Agriculture Action Plan for China's Agenda 21*, China Agriculture Press, Beijing, China.

Ministry of Forestry, 1978, *National Forest Resource Statistics (1973-1976)*, Beijing, China.

Ministry of Forestry, 1983, *National Forest Resource Statistics (1977-1981)*, Beijing, China.

Ministry of Forestry, 1990, *National Forest Resource Statistics (1984-1988)*, Beijing, China.

Ministry of Forestry, 1996, *National Forest Resource Statistics (1989-1993)*, Beijing, China.

Nickum, J.E., 1982, *Irrigation Management in China: A Review of the Literature*, The World Bank, Washington, DC, USA.

Nilsson, S., 1999, *The Forest Sector of China*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Parikh, A., and Thorbecke, E., 1996, Impact of rural industrialization on village life and economy: A social accounting matrix approach, *Economic Development and Cultural Change*, **44**(2):351–377.

Price Statistical Yearbook of China, 1992, China Statistics Press, Beijing, China.

Richardson, S.D., 1990, *Forests and Forestry in China: Changing Patterns of Resource Development*, Island Press, Washington, DC and Covelo, CA, USA.

Rozelle, S., Benzinger, V., Li, G., and Huang, J., 1998, Forest Exploitation and Protection in Reform China: Assessing the Impact of Policy, Tenure, and Economic Growth, unpublished manuscript.

Rozelle, V., and Huang, J., 1997, The impact of environmental degradation on grain production in China: 1975–1990, *Economic Geography*, **73**(1):44–66.

Shen, J., and Spence, N.A., 1996, Modeling urban-rural population growth in China, *Environment and Planning A*, **28**:1417–1444.

Shi, P., and Xu, J., 2000, Deforestation in China, unpublished manuscript.

Spear, A., He, S., and Nailer, C., 1997, *China's Infrastructure: Sectoral and Financing Outlook*, East Asia Analytical Unit, Canberra, Australia.

State Statistical Bureau of China, 1996, *1992 Input-Output Table of China (in Value)*, China Statistical Publishing House, Beijing, China.

State Statistical Bureau of China, 1997, *1992 Regional Input-Output Table of China (in Value)*, China Statistical Publishing House, Department of Economic Accounting, Beijing, China.

Subramanian, S., and Sadoulet, E., 1990, The transmission of production fluctuations and technical change in a village economy: A social accounting matrix approach, *Economic Development and Cultural Change*, **39**(1):131–175.

Sun, L., 2000, *The Specification of the LUC Core Model for GAMS Implementation*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Sun, L., and Li, X., 1997, *Driving Forces of Arable Land Conversion in China*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

TEI (Technical and Economic Institute of the State Planning Commission), 1994, The development strategy for cars to be used by households in China, *Economic Daily*, October 24.

Tian, X., 1999, Market orientation and regional economic disparities in China, *Post-Communist Economies*, **11**(2):161–172.

United Nations Population Division, 1998, *World's Population Prospects 1950–2050: The 1998 Revision*, United Nations, New York, NY, USA.

World Bank, 1985, *China: Agriculture to the Year 2000*, The World Bank, Washington, DC, USA.

World Bank, 1992, *Reform and the Role of the Plan in the 1990s*, The World Bank, Washington, DC, USA.

World Bank, 1997a, *At China's Table: Food Security Options*, The World Bank, Washington, DC, USA.

World Bank, 1997b, *China 2020: Development Challenges in the New Century*, The World Bank, Washington, DC, USA.

Wu, H.X., and Findlay, C., 1997, China's grain demand and supply: Trade implications, in *Agricultural Policies in China*, Organisation for Economic Co-operation and Development, Paris, France.

Xie, J., and Jia, K., 1994, Assessment of crop production potential: Potential and potential population supporting capacity of land, in *Land Resources, Use and Productivity Assessment in China*, State Land Administration of People's Republic of China, Food and Agriculture Organization of the United Nations, The Program Working Group of The Population-Bearing Capacity of Land in China, Beijing, China.

Yin, R., 1994, China's rural forestry since 1949, *Journal of World Forest Resource Management*, **7**:73–100.

Zheng, Z., and Tang, C., 1994, Assessment of livestock production potential, in *Land Resources, Use and Productivity Assessment in China*, State Land Administration of People's Republic of China, Food and Agriculture Organization of the United Nations, The Program Working Group of The Population-bearing Capacity of Land in China, Beijing, China.

Chapter 13

Development and Operationalization of an Integrated Modeling Approach for Land Use Planning and Policy Analysis

AbuBakr AbdelAziz Mohamed

Abstract

Despite the recognition of its importance in natural resources management, the problem of integrating agro-ecological and agro-economic analysis remains a major challenge and has as yet not been solved in a satisfactory manner. This chapter contributes to development and operationalization of a model for land-use planning and policy analysis that integrates agro-ecological and agro-economic information in such a way that land-use policy options at the sub-regional level can be formulated and evaluated with the aim of aiding policy makers. The chapter consists of five sections. Section 13.1 gives an introduction to an integrated agro-economic and agro-ecological modeling approach for land-use planning and policy analysis. Main conceptual and methodological challenges that stand in the way of integration are analyzed and described in Section 13.2. These challenges form the basis for the

development and operationalization of the integrated model for land-use planning and policy analysis (Section 13.3). The performance of the approach is discussed in Section 13.4. Finally, a discussion and conclusions on strengths and weaknesses of the approach is presented in Section 13.5.

Keywords: Land-use planning and policy analysis, farm aggregation and mapping, integrated land-use system, land-use policy scenarios, GIS, LP, multi-criteria evaluation.

13.1 Introduction

Land-use decisions are rooted in the physical and biological sciences, but they are driven by human behavior. Hence, resource problems generally have at least an agro-economic and an agro-ecological dimension, but when land-use decisions are modeled, the focus is very often on *either* of the two. However, the complexity and range of dimensions in land-use problems are calling for an integrated approach. Despite the recognition of its importance in land-use planning and policy analysis, integration of agro-economic and agro-ecological analysis has not been achieved in a satisfactory manner, and remains a major challenge (van Latesteijn, 1999; van Ittersum, 1998).

Development of an integrated modeling approach to land-use planning and policy analysis appears specifically hampered by lack of an adequate research methodology (RAWOO, 1989). However, the integrated approach is essential, if scientific research is to make an effective contribution to protecting and restoring natural resources. This chapter aims to contribute to development of an integrated approach for modeling land-use planning and policy analysis, one that may serve as a decision support tool for policy makers to formulate and evaluate land-use policy scenarios at the sub-regional level.

13.2 The Integration Problem

Although, in theory, the need for integration of agro-economic and agro-ecological analysis in land-use planning and policy analysis is well recognized (see, for example, Fresco *et al*., 1992; Stomph *et al*., 1994; Alfaro *et al*., 1994), applications are still hampered by major obstacles that render difficult the integration process. In the realm of agricultural planning, many of these conceptual and methodological constraints to integration have been discussed (see, for example, Malingreau and Mangunsukardjo, 1978; Luning 1986; Braat and Van Lierop, 1987; RAWOO, 1989; Van Diepen *et al*., 1991; Fresco *et al*., 1992; Hengsdijk and Kruseman, 1993;

Sharifi and Van Keulen, 1994; Stomph *et al*., 1994; Pichett *et al*., 1994; Schipper, 1996). From these reviews the main constraints have been distilled.

13.2.1 Difficulty of integrating disciplines

From a methodological point of view, integrating agro-ecological and agro-economic information represents a major challenge because the two types of information rely on different disciplines or paradigms. For more effective integration of these disciplines, an integrated interdisciplinary unit of analysis is required. Mohamed and Sharifi (1998) identified the following major challenges to creating such an integrated spatial unit. They include the different nature and focus of disciplines, different units of analysis, different hierarchical levels of analysis, and difficulty of spatial linking of disciplines. The difference in nature and focus of disciplines is reflected in the observation (Luning, 1986), that scientists from different disciplines think, understand, and approach the “same phenomenon or the same problem” in different ways, which leads to different units of analysis and perhaps to different criteria for assessment.

Land use from an agro-ecological point of view is described in terms of land units that can be used to discriminate among alternative land uses (FAO, 1976). From an agro-economic point of view, often the guiding principle for land-use decision making is linked to the aspirations of farm households. The challenge is, therefore, to combine the land unit and the farm household into one integrated spatial unit. Most of the information on agro-ecological aspects can easily be geo-referenced or mapped, which is difficult for information on agro-economic aspects, as that is descriptive or conceptual and not geo-referenced. This lack of geo-referencing for agro-economic information frustrates identification of an integrated spatial unit.

The discussion on the difference in units of analysis is closely linked to the hierarchical levels at which both disciplines operate or exist. The systems hierarchy proposed in agriculture by, for example, Van Dyne and Abramsky (1975), Fresco (1986), Conway (1987), and Fresco *et al*. (1992), is derived from the application of the hierarchical structure of ecology to agriculture. Stomph *et al*. (1994) has pointed to one major disadvantage of the suggested hierarchical approach in agriculture: at the lower levels, mainly biophysical criteria are used for classification, and at higher levels mainly socioeconomic or administrative criteria are deployed. The challenge is, therefore, to identify a level in the hierarchy of systems at which both realms meet.

13.2.2 Difficulty of integration levels

Integration of agro-economic and agro-ecological aspects for land-use planning and policy analysis requires combining data from different spatial levels. Linking levels of analysis, therefore, is an important prerequisite for the integration (Fresco *et al*., 1992). Land-use decisions involve choices at two spatial levels, at least. At the regional level, a policy maker is trying to decide how best to allocate limited resources in the face of uncertainty about all its allocation consequences. This uncertainty derives from the uncertainty about farmers' response to policy changes. At the farm level, farmers face the problem of how best to respond to the new policy environment, given their own resources, objectives, and limitations of actions (Hazell and Norton, 1986). To solve the macro-level decision problem, the uncertainty about farmers' responses has to be reduced, which ideally is achieved by aggregating the behavior of individual farms.

In agricultural planning, scaling up the analysis from farm level to regional level is the source of the aggregation problem. In this transition, aggregation bias is introduced, because not all farms are similar (Jansen and Stoorvogel, 1998). In land-use planning, the aggregation from farm level to regional level of analysis remains a pressing and unsolved issue, that requires much further research (Fresco *et al.*, 1992). The objective is to avoid or minimize aggregation bias, when linking farm-level information to regional level analysis.

13.2.3 Insufficient attention to quantifying socioeconomic analysis

Integrating biophysical and socioeconomic aspects of land-use practices requires a format for quantitative description of both. Unfortunately, there is a large discrepancy between the detailed quantitative description of the biophysical aspects, and the broad qualitative terms in which the socioeconomic aspects are generally described. While tremendous progress has been made in the quantification of the biophysical aspects, similar descriptions of the socioeconomic aspects are still in their infancy (Stomph and Fresco, 1991; Van Rheenen, 1995; Stomph *et al*., 1994; Schipper, 1996). An excellent example of the insufficient attention given to socioeconomic analysis of land-use practices is the economic critique on land evaluation by Schipper (1996). An equally illustrative example of the qualitative nature of socioeconomic analysis of land-use practices is the argument by Van Rheenen (1995), that farming systems analysis methods are too qualitative when used to assess policy making. One of the challenges for the integration is, therefore, quantification of the socioeconomic component as a part of an integrated land-use system analysis.

13.2.4 Difficulty of incorporating diversity of stakeholders

Land-use planning in its simplest form is the allocation of land to various categories of use according to predetermined criteria (Van Diepen *et al*., 1991). An important step in land-use planning is, therefore, the selection of the preferred land-use type for a certain land unit. Because land units may be suitable for more than one land-use type, choices must often be made. Normally, selection of the "best" land-use types for the land units of a region, district, or village must take into account a number of goals (Huizing and Bronsveld, 1994).

Land-use planning, therefore, deals with multipurpose use of land, trade-offs among different functions of the land, and conflicting interests among the different categories of stakeholders and between collective and individual goals and needs (Van Diepen *et al*., 1991). When only one goal has to be pursued (optimized) the approach is straightforward. However, with a number of possibly conflicting goals, the choice for a certain development path becomes dependent on the relative weight attached to each of the goals, which is not necessarily the same for various interest groups (Fresco *et al*., 1992). One of the challenging issues, therefore, is the integration of this diversity of stakeholders and their goals into the land-use planning process (Fresco, 1994).

13.3 Conceptual Framework and Building Blocks of the Integrated Approach

The integrated modeling approach for land-use planning and policy analysis, presented here, derives its conceptual foundation largely from an adaptation of the theory of economic policy of agricultural sector analysis to land-use planning. Agricultural sector analysis offers great potential as a tool for planning and policy analysis (Thorbecke and Hall, 1990; Hazell and Norton, 1986) and has proven its usefulness in land-use analysis (Hengsdijk and Kruseman, 1993; Schipper *et al*., 1995; and Schipper, 1996). Moreover, the theory of economic policy of sector analysis provides the common denominator and language to contributions from various disciplines involved in agriculture (Thorbecke and Hall, 1990). The approach is structured in a set of interrelated blocks, each containing a number of steps, and requiring a number of tools and/or methods for its operationalization. The structure of the basic conceptual framework of the integrated approach and its main building blocks are shown in *Figure 13.1*.

13.3.1 Designation of integrated spatial units for land-use modeling

Modeling on the basis of land units in land-use planning and policy analysis may result in biophysically, but not necessarily agro-economically, homogeneous units,

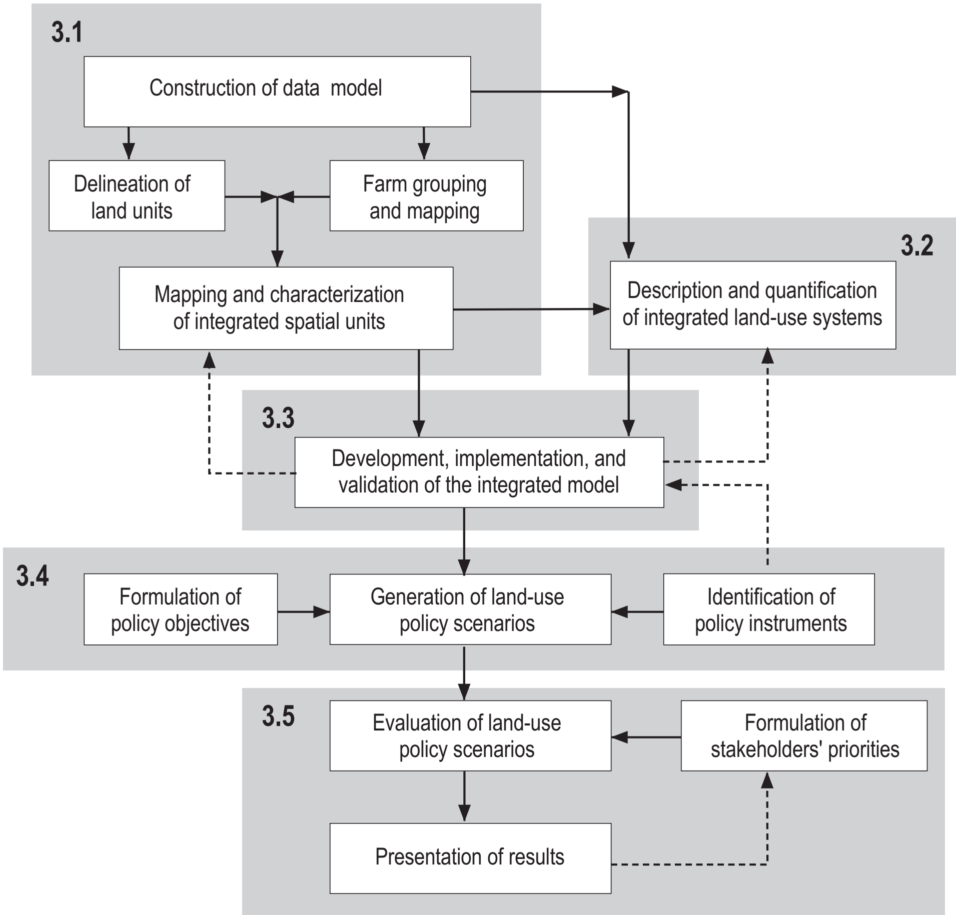


Figure 13.1. Basic conceptual framework of the integrated approach for land-use planning and policy analysis and its main building blocks.

whereas modeling on the basis of farm types may result in agro-economically, but not necessarily biophysically, homogeneous units. To define a more integrated unit, the concept of "farm-type land unit" (FTLU) or simply integrated unit (IU) is developed (Mohamed, 2000). To illustrate the concept of FTLU, consider a region with two land units: LU1 and LU2 and two farm types: FT1 and FT2 as depicted in *Figure 13.2*.

The concept of FTLU implicitly assumes that both FTs and LUs can be mapped. While LUs can easily be geo-referenced and presented on a map, information on FTs is generally difficult to map. Without geo-referencing of FTs, it is difficult, if not impossible to link agro-economic and biophysical disciplines. The concept presented here has been operationalized in a four-step procedure: (1) a simple

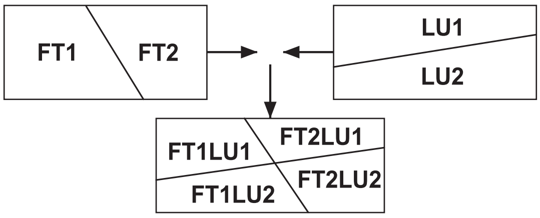


Figure 13.2. Schematic presentation of the concept "Farm Type Land Unit (FTLU)."

data model has been constructed; (2) a procedure for farm grouping and mapping has been developed and implemented; (3) land units have been delineated; and (4) mapped farm types and delineated land units are integrated in farm-type land units (FTLUs).

Construction of a Data Model

The approach is illustrated for Amol sub-region, Mazandaran Province, Iran. The data for the case study have been derived from various sources, e.g., detailed results of the Agricultural Census from the Iran Statistical Centre, farm survey data from the Agricultural Statistics and Information Department of the Ministry of Agriculture, and data from regional and sub-regional offices.

The available data on farming systems contains information on 18,662 farm households located in 277 villages. For each farm household details are given on 140 attributes. Such a large amount of data items will be of little use unless the data is structured in a meaningful way (Howe, 1989). Therefore, a simple database model has been designed so that unnecessary duplication of the data is minimized and so that the data can be quickly retrieved in all required sequences (Benyon, 1990). Moreover, the database makes it possible to perform tasks that involve handling large amounts of data (Date, 1990).

Hence, the available farming systems information has been organized and stored in a GIS database using the Arc/Info software package. In this database, spatial data and thematic attributes that are linked by a unique identifier (village code) characterize each entity or feature (village in this case). Spatial data comprise location information, describing X and Y coordinates (longitude and latitude) of each village. Thematic attributes include the agro-economic characteristics of farm systems.

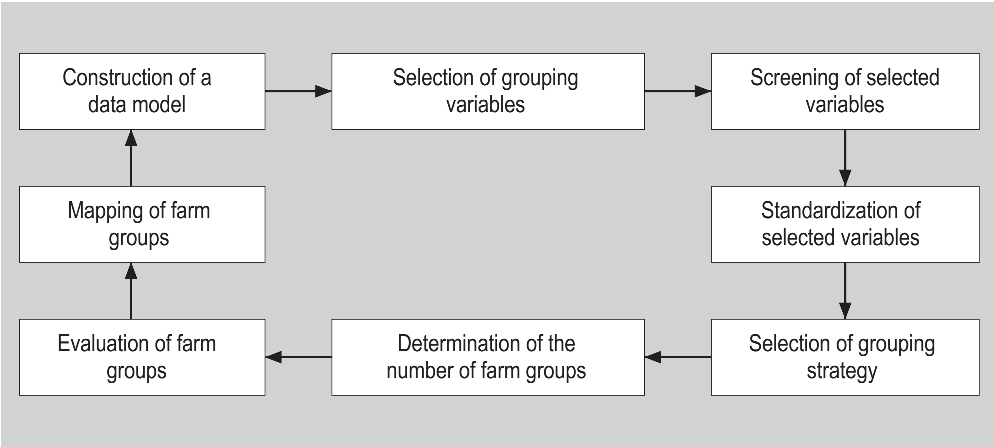


Figure 13.3. Skeleton of the farm grouping and mapping methodology.

Farm Grouping and Mapping

An important step in the operationalization of the concept of farm-type land unit is the procedure for farm grouping and mapping. The skeleton of the farm grouping and mapping methodology includes eight basic steps (*Figure 13.3*). Each of the steps in the methodology contains a number of sub-steps, and requires various methods and/or techniques for its operationalization. The process of farm grouping and mapping is considered an endless loop, in which new insights are obtained and new ideas generated during each pass. The feedback chain is the basis for learning how to do better and improve the grouping. The end result consists of farm types or groups that match the purpose reasonably well.

Because the main concern for farm grouping, in this study, is to eliminate aggregation bias while integrating levels of analysis, this procedure begins by exploring the requirements for bias-free aggregation (conceptual analysis), and then subsequently investigates the criteria used for farm-systems grouping in some empirical studies (empirical analysis). Both conceptual and empirical analyses are then combined to recommend operational variables for grouping (see *Table 13.1*). Aggregation bias can only be avoided if farming systems are classified into groups, which are defined according to the theoretical requirements of Day (1963).

The requirements set forth in *Table 13.1* (institutional, technological, and pecunious similarity) are used as guidelines for the selection of grouping criteria. Moreover, to facilitate integrating the resultant farm systems with land units, the mapping of farm systems is of prime importance. It is made possible by incorporating location attributes into the grouping criteria. That is why the important

Table 13.1. Framework of the proposed criteria for farm grouping.

Grouping criterion/variable	Acronym	Unit of measurement
Institutional similarity		
Land area per farm household	LNA_FHH	Proportion to largest (%)
Land area under irrigated farming	LNA_IRF	Proportion to total land area (%)
Land area under dry farming	LN_DRF	Proportion to total land area (%)
Land area exploited under private tenancy	LNA_PRV	Proportion to total land area (%)
Land area exploited under partnership tenancy	LNA_PRN	Proportion to total land area (%)
Farm households with farm size $<$ 1 ha	FHH1	Proportion to total farm households (%)
Farm households with farm size between 1 and 3 ha	FHH1_3	Proportion to total farm households (%)
Farm households with farm size $>$ 3 ha	FHH3	Proportion to total farm households (%)
Ground water availability	GWA	m^3/ha
Ground water pumping capacity	GWPC	hp/ha
Family labor availability	FLBA	mnd/ha
Technological similarity		
Quantity of urea applied	FRT_N	kg/ha
Quantity of phosphate applied	FRT_P	kg/ha
Mechanical power availability for tillage	MPA _TL	hp/ha
Mechanical power availability for rice threshing	MPA_TH	hp/ha
Pecunious similarity		
Overall production efficiency	OPE	Proportion to highest (%)
Location proximity		
Geographical longitudinal co-ordinate	LON_DM	Degree and minute
Geographical latitudinal co-ordinate	LAT_DM	Degree and minute

requirement called "location proximity" has been added to the proposed variables (see *Table 13.1*).

To determine if a justification exists for grouping, to prevent the inappropriate application of a clustering strategy, and to provide information on the fundamental nature of the data, correlation analysis and coefficient of variation are used to screen the selected variables. Certain variables have also been standardized to remove the arbitrary effects of measuring variables in different units and to make variables contribute more equally to the similarities among farming systems. This step consists of specifying theoretically possible proximity measures and clustering methods, comparing the performance of the methods in some empirical work, and then combining the conceptual and empirical analysis to select grouping strategies.

Ten possible grouping strategies were applied and the result that fits the purpose well was taken. The 10 strategies combine three proximity measures – squared Euclidean (SQ); city block (CB); and cosine (CS) – and four clustering methods – average linkage within groups (WG); average linkage between groups (BG); complete linkage (CL); and ward (WD). Several methods for determining the optimal number of farm types were used and all the results were analyzed. These methods included informal ones that used the dendrogram and the agglomerative schedule, and formal ones such as the cluster indicator value (CIV).

The evaluation procedure was a step-by-step scrutinizing of the set of grouping strategies to find the one with the most distinct clusters. It consisted of four steps. It started by comparing the agreement between the solutions produced by the ten grouping strategies using the rand index. Only those groupings that produce similar results or agree closely were carried onto further validation. Farm clusters produced by these groupings were tested for their significance of difference using Kruskal-Walis H statistics. Only those groupings whose clusters were significantly different were ranked and the one with the most distinct clusters was selected. Then its clusters were described and characterized with their particular features.

To facilitate mapping farm types, a (partial) link was established between the geographical information system (GIS) and grouping models. This link is based on the general framework for the GIS-model link suggested by Stoorvogel (1995). GIS supplies input data for the grouping models and accepts modeling results for further processing, analysis, and presentation. Mapping of farm groups has been done in a four-step procedure.

In Step 1, the identifier (village code) and its corresponding agro-economic and location attributes were organized in a table and exported from GIS to an Excel spreadsheet. On the basis of the procedures developed for the selection and screening of variables used in farm grouping, a number of mathematical and statistical operations were performed in such a way that attributes exported from the GIS database are converted in specific input parameters required for farm grouping.

In Step 2, and on the basis of the methodology developed for farm grouping and mapping, grouping model runs were carried out using 10 alternative grouping strategies. During the grouping runs, the identifiers to the original GIS database file were preserved, to link the grouping results to the base map in GIS. Cluster memberships at specified cluster levels of the alternative groupings were saved as new variables in the active SPSS file. In other words, for each village, a value indicating the cluster to which the village belongs in a given solution, is stored in a specific variable name. For example, a new variable CSCL indicates the cluster to which each village belongs when four clusters are produced using the grouping strategy that combines a proximity measure cosine (CS) and a clustering method complete linkage (CL). These new variables were used in subsequent analyses for validating alternative groupings, and for testing the significance and contiguity of farm types produced by these groupings. On the basis of these analyses, the alternative groupings were compared, tested, and ranked.

In Step 3, cluster memberships, stored as new variables in the active SPSS file, were saved as an SPSS output file which in turn was converted to text file format. The text file was loaded directly into Arc View GIS as a table. Then the tabular data were added to the base map by joining it to the attribute table of its theme. Saving the project containing the joint procedure[1] saved the definition of the joint. By joining, all the fields from the tabular data were appended to the attributes of the base map. In this way, Arc View GIS was used to visualize results of the alternative groupings, to check whether they form contiguous farm units.

The statistical significance of farm clusters produced by the various classification strategies was tested using Kruskal–Wallis H statistics. The test was simply applied to each of the grouping strategies in turn. Once they have passed the statistical significance test, these grouping strategies must be compared to select the one that produces distinct and contiguous farm types. The selection was based on the level of statistical significance, and on the contiguity of farm types, using a two-step procedure. In the first step, the value of Kruskal–Wallis H statistics was used to establish which grouping strategy was likely to produce the most distinct farm types. This decision was based on transformation of the H statistics for each variable to values in the range from 0 to 1. Grouping strategies were then ranked on the basis of the mean of the values of all these variables. In the second step, the grouping strategies were visualized using Arcview GIS to examine whether they also result in formation of contiguous farm types. Then results of the two-step procedure were compared to select a grouping strategy with distinct and contiguous farm types. As CSCL produces the most distinct and contiguous farm types, this grouping strategy was selected.

[1]The joining is based on a common field “the identifier: village code” that is part of both tables.

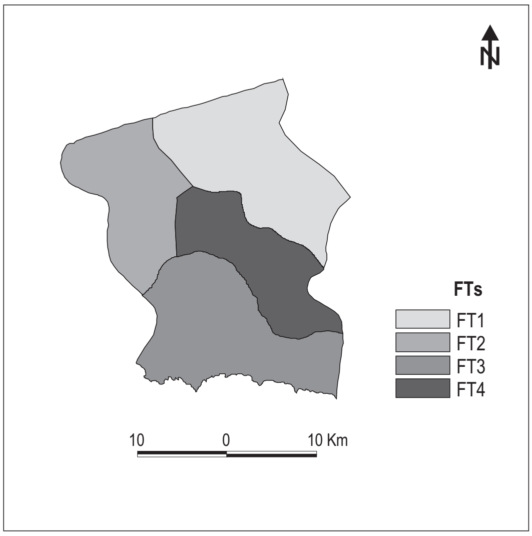


Figure 13.4. Mapping farm types (FTs).

Step 4 consists of delineating the farm types produced by the selected grouping strategy CSCL. In this step, new polygon features are created by merging point features (villages) that have the same value for cluster membership and that clustered in spatially contiguous units into one polygon using the Arc-Edit module of Arc/Info. Merging was performed on screen using a mouse. The base map, which includes cluster memberships as new attributes, is used as background coverage and the farm types are displayed in different colors to show their boundaries. Arcs have been drawn to approximate these boundaries. The merging of villages, clustered in the same farm type and forming a spatially contiguous unit in one polygon (*Figure 13.4*), creates a new map. Following this procedure, it is necessary to reclassify some villages that fall inside a specific farm unit, but belong to a different farm type. Regrouping of these villages was performed by changing their cluster membership in the SPSS file created in Step 3. Then, statistical testing was carried out again to assure the significance of these clusters.

Delineation of Land Units (LUs)

For delineation of land units, the aim is to make use of existing information from previous studies carried out in the sub-region. Land units are defined, in this study, as a combination of agro-climatic zones and soil series. The procedure for mapping land units consists of three steps. In the first step, agro-climatic zones (ACZs) are delineated; then soil series (SSs) are mapped in the second step; and finally, in the third step, land units are identified by overlaying the two coverages.

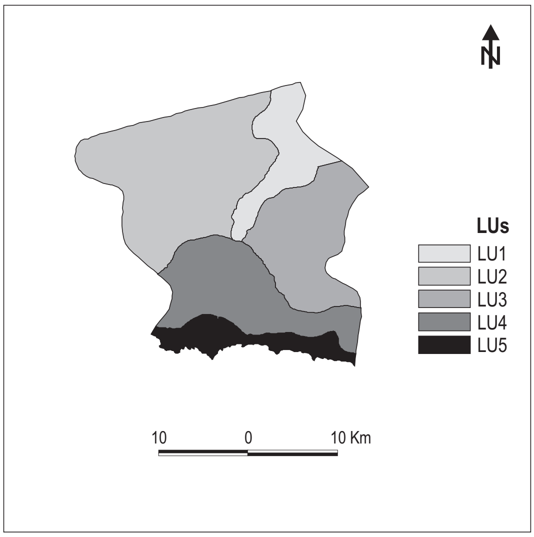


Figure 13.5. Delineation of land units (LUs).

Recent activities to inventory and analyze the climatic resource in the region have been carried out jointly by the Ministry of Agriculture of the Islamic Republic of Iran and FAO, using the FAO-Agro-Climatic Zoning methodology and procedures. These activities are documented in Taazimi (1995). The purpose of these climatic resource inventories is to provide the necessary information for analyzing the production potential of agricultural crops in the region. On the basis of these technical reports, three climatic zones have been distinguished in the study area. The soils of the region have been studied in many surveys, part of which have been collated and correlated by King (1995), for development of a grouping of soil series. On the basis of this report, the soils of the study area are classified into five series.

Two coverages, comprising the geographical distribution of soil series (SSs) and the agro-climatic zones (ACZs), were stored in GIS together with attribute data. A map overlay was implemented using Arc/Info software, which resulted in a map with new units (*Figure 13.5*). These units, characterized by similar soils and climates, are referred to as land units (LUs). These LUs form the bio-physical component of the integrated unit.

Identification of the Integrated Units

Having mapped farm types (FTs) and identified land units (LUs), the next step was to integrate these two disciplinary units by spatially linking them. The link was established in a GIS environment through a map overlay procedure. Two maps, the

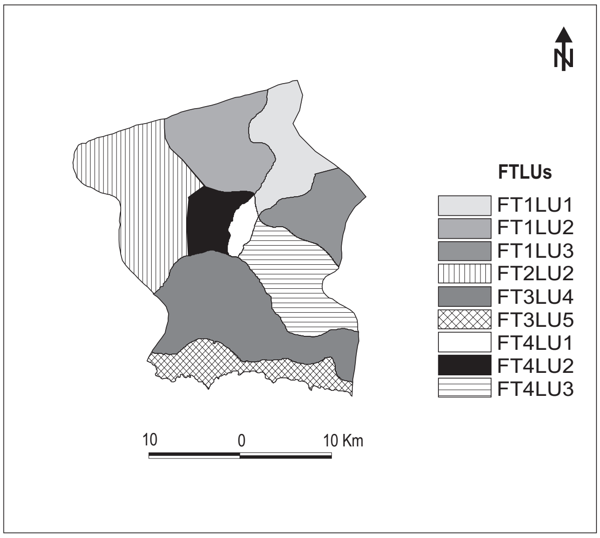


Figure 13.6. Identification of integrated units (IUs).

map of farm types and the land units map, were combined in an overlay procedure to establish the location of farm types in the various land units. Each unique combination of FT and LU is referred to as a farm-type land unit (FTLU) or, alternatively, an integrated unit (*Figure 13.6*). This integrated unit forms the basis for definition, description, and analysis of land-use systems.

13.3.2 Description and quantification of integrated land-use systems

The term land-use system (LUS) can be used for any description of land use at land-unit level. Land-use systems have been defined differently in various studies, depending largely on their purpose: Beek (1978), FAO (1976; 1983; 1984; 1985; 1987), Driessen and Konijn (1992), De Koning *et al*. (1992), Van Lanen (1991), Van Duivenbooden *et al*. (1991), Jansen and Schipper (1995), Schipper (1996), and Zonneveld (1997). However, all these definitions have in common, that land units are defined on the basis of specified biophysical characteristics only. Purely socioeconomic characteristics are not included in the concept of land. This creates the necessity of introducing the socioeconomic specifications, when included in the description of land-use types, in an operational way in land-use planning and policy analysis.

Although theoretically many definitions recognize that land-use types are parts of farm systems, and therefore not independent, in practice, the suitability of land units for specific land-use types is assessed in isolation, without taking into account the context of the farm as a unit of decision making. In a way, this implies looking at land use at a (sub-) regional level, omitting the farm level. Many land-use system assessments, although relevant in themselves, are therefore less applicable for land-use planning and policy analysis, and certainly less applicable as a basis for implementing a proposed land-use change (Polman *et al*., 1982; Fresco *et al*., 1992; Erenstein and Schipper, 1993). To deal with the observed omission in the definition of the land-use system, the concept of integrated land-use system (ILUS) has been introduced (Mohamed, 2000).

This concept is based on the logical argument of Stomph *et al*. (1994) that land-use systems, irrespective of the level at which they are defined, are integrated systems and their description should include both biophysical and socioeconomic characteristics. Only then can a comparison can be made between what land can supply and the demands on land use. In accordance with the definition of systems (Fresco, 1986), inputs and outputs are defined, and the transformation processes from inputs to outputs in the system are identified and quantitatively described.

The simplified diagram in *Figure 13.7* illustrates some of the important components considered. ILUS itself is not a closed system but a sub-system of a larger system at a higher level of aggregation. The concept ILUS is proposed for a specific form of describing a land-use system, defined as a combination of a farm-type land unit (FTLU), a land-use type (LUT), and a production technique. In the present chapter, land-use types are described in relation to farm-type land units. Any land-use type can be practiced in various socioeconomic and biophysical settings, depending on farm-type land unit. Various (agronomic and socioeconomic) technical specifications can be defined for a given land-use type, dictated by its biophysical and socioeconomic settings. Combining information on the settings and specifications with information on type of land use (e.g., crop commodity) allows specification of land-use types with both biophysical requirements and socioeconomic requirements.

Land-use types, as components of integrated land-use systems, are described in terms of agronomic technical specifications and operation sequences (Stomph *et al*., 1994; Jansen and Schipper, 1995; Schipper 1996). Within the integrated land-use system, most operations have to be carried out in a given order or sequence, determined by the growth and development pattern of a particular crop. In this chapter, that order has been maintained in describing land-use types as parts of land-use systems, in terms of the operations involved (Van Heemst *et al*., 1981; Van Heemst, 1986): land preparation, preparation of plant material, planting/seeding, fertilizer application, irrigation, weeding and thinning, biocide application, harvesting, and threshing.

The following attributes fully define any operational sequence: timing of operations; types and quantities of applied material inputs; types of implements; type

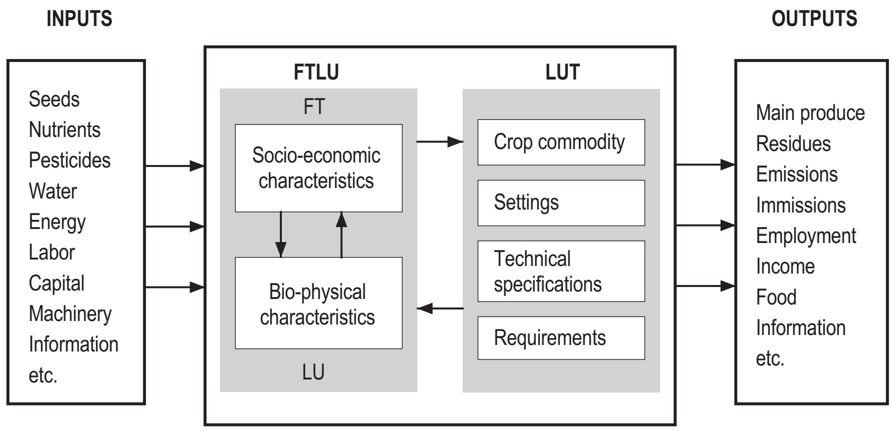


Figure 13.7. Simplified diagram of an integrated land-use system.

and quantity of traction power source; and types of outputs exported from the system (Stomph *et al*., 1994). Operation sequences in this chapter have been identified on the basis of farm surveys, agricultural statistics and reports, agricultural census, expert knowledge, and results of theoretical and empirical studies in similar or other regions. For each of the operations, various methods or techniques may be used, depending on farm-type land unit and crop commodity, distinguished by type of traction, equipment and materials, inputs of labor and materials, and timing of the operation. Generally, three major types of alternative techniques can be distinguished. The first criterion refers to the timing of the operation, the second to the amount of non-factor input per unit area (e.g., amount of fertilizer per hectare), the third to factor substitution relations, such as when choosing among different levels of mechanization.

Combining information contained in operational sequences with information on integrated land-use systems allows description of land-use systems in terms of these operational sequences. Such a description then serves as a basis for calculation of the required input-output coefficients. This has the advantage, that land-use systems do not have to be described again for each change in the calculation of the coefficients. Each unique operational sequence within an ILUS can be interpreted as a specific (land-use) activity. Each activity is defined and described quantitatively in terms of inputs for production and outputs, desired as well as undesired.

Both current and alternative land-use systems are incorporated in the analysis. Input and output coefficients of current land-use systems have been derived from information from the detailed farm survey. To specify a sufficiently wide range of land-use options, in addition to those actually observed for a given farm-type land unit, alternative production techniques are specified. Those represent technically feasible production techniques that are not yet practiced by farmers in the region, and that aim at maintaining the resource base and protecting the environment. For quantification of alternative land-use systems, a so-called target oriented approach is applied (van Ittersum and Rabbinge, 1997), in which the combination of inputs required to realize a specific level of outputs is estimated based on knowledge of the underlying biophysical processes.

All inputs and outputs of an ILUS activity are expressed as physical quantities or monetary values or time or power per hectare. This analysis provides quantitative information for each of the considered combinations of land-use system and operation sequences on material (seeds, fertilizers, pesticides, etc.); water and labor input requirements and their distribution over the year; time and power needed in terms of implements and traction sources other than human; costs of operation sequences in terms of material, water and labor inputs and machinery services and their distribution over the year; amounts and prices of harvested products; and amounts of other outputs exported from the system.

13.3.3 Development of the integrated model

A linear programming model for land-use planning and policy analysis is developed, and used as a tool for integrating socioeconomic and biophysical components of land-use systems in such a way that land-use policy options at the sub-regional level can be generated. The purpose of the model is to analyze the possible effects of policy measures on farm household land-use decisions and their consequences for realization of regional agricultural development policy objectives. For a better understanding of the effectiveness of different policy measures on agricultural development, a micro-oriented analysis of farm-level response is indispensable (Van Keulen *et al*., 1998).

Hence, the model developed in this chapter departs from the farm level. The model integrates different levels of aggregation. The model is based on integrated land-use systems (ILUSs) as core units at the activity level. The first level of aggregation is ILUS at the farm-type land unit (FTLU) level, an aggregation level below both farm type (FT) and land unit (LU). Therefore, FTLUs can be aggregated to LU level, over the various FTs, or to FT level, over the various LUs. In this way, aggregation of FTLUs yields either land units with strong socioeconomic components, or farm types with strong biophysical components. And finally, FTs or LUs can be aggregated to the sector at the sub-regional level.

For simulating the region's response to possible policy changes, a positive (or descriptive) linear programming model has been used, in which (policy) goal variables do not enter the objective function, nor are their levels constrained (Hazell and Norton, 1986). To include goal variables in the objective function would imply losing the simulating (positive) role of the model. The same comments apply to constraining the levels of goal variables. However, maximizing a policy goal variable directly can serve an analytical purpose, i.e., to establish the frontier, or its maximum level that conceivably would be attainable. But even this limit may not be very useful, because it may show little relation to the points that are attainable under the market systems. These considerations reinforce the arguments that, to address policy issues use of a positive model, via a sequence of experiments, involving changes in policy parameters is preferable to using a frontier.

The objective function of the model is the sum of net benefits over all farm-type land units. This objective function calculates all benefits and costs from the farmers' point of view, thus mimicking a postulated objective of farmers and thereby introducing an aspect of their behavior. The objective function is maximized subject to a number of conditions. Except for the objective function, all 99 groups of model equations can be classified according to the dichotomy of balances and constraints. Equations include resource constraints and other restrictions, product balances, input balances, pricing and costing balances, and sustainability balances.

The model includes a wide range of production activities, representing different crops, different ways of producing them, and cultivation on different farm-type land units. To create sufficient flexibility, various sub-land-use types are defined for the same land-use type, differentiated according to production technique, and incorporated in the model as different types of land-use activities. The wide range of land-use activities included provides a wide scope for selection of appropriate technique (Hazell and Norton, 1986).

In the model, production units (or resource supply sources) are farm-type land units. Resource endowments of farmers in each of the farm-type land units include land, irrigation water, family labor, and farm machinery. Separation of resource supply sources is the basic rule under which the model has been specified, i.e., they are differentiated in space, by resource type, and over time. The land resource is differentiated as irrigated and non-irrigated, when relevant. Irrigation water supply is differentiated by supply source, and labor supply into family and hired labor. Timing of operations within production activities is crucial. If seasonal patterns of resource availability are ignored, the solution may be unrealistic by showing surplus of a resource, when in fact seasonal resource shortages prohibit realization of specified production levels. Hence, land, irrigation water, and labor resource availability have been specified on a monthly basis. Time thus enters the model as a characteristic of resource inputs. Introducing this seasonality may constrain the model solution, and thus lead to lower values of the objective function.

It is therefore important to also consider any options the farmer has for reducing seasonal bottlenecks in resource availability (Hazell and Norton, 1986). Complementary resource supplies at different aggregation levels have been specified: family labor supply, for example, is constrained at the farm-type land unit level, whereas hired labor supply is constrained at the sub-regional level. In the model, resources are supplied through a separate set of variables (called activities or columns), and balance equations are incorporated to ensure equilibrium on factor markets. Resource supply is combined with resource constraints by introducing resource balance equations, in which the demand for the resource by land-use activities is balanced by the supply of the resource per farm-type land unit, resource type, and month.

A competitive market form is assumed, implying that no producer has a sufficiently large scale of operations to influence the market price (Hazell and Norton, 1986). This description of the market mechanism closely represents the actual processes that determine production in the Amol sub-region. When used in this sense, a competitive market does not imply absence of market imperfections. Market imperfections have been incorporated in the model through exogenous spatial price differentials, as observed farm-gate product prices in different farm-type land units, as obtained from the farm survey data. This description provides a realistic picture of the market conditions faced by farmers.

Land-use systems are characterized by output coefficients and input coefficients. Therefore, land-use activities in the model comprise input demand activities and product supply activities, i.e., two market functions: product supply and input demand. These functions are part of the model's structure incorporated through definition of appropriate variables and balance equations to calculate total input demand and output supply. In the balance equations, these variables are set equal to the values of land-use activities multiplied by the coefficients representing the relevant demand or supply quantities.

Product supply is calculated through commodity balances, specified per product type, for the sub-region as a whole, as well as for any particular farm type or farm-type land unit within the sub-region. Also input demand is calculated per input type at monthly intervals or on an annual basis (depending on input type) and at different levels of aggregation: farm-type land unit, farm type, or sub-regional. The demand for land, irrigation water, and labor is defined on a monthly basis. All other inputs are considered on an annual basis, including services of farm machinery.

Unlike input demand functions, which are implicit in the land-use activities, input supply functions are explicitly specified in the model structure. The supply functions for many factors are simple: either perfectly inelastic or perfectly elastic. For land the supply function typically is perfectly inelastic in the short run, as is that for irrigation water. For purchased inputs, the supply functions are perfectly elastic at the given price. However, for some factors supply functions fall between these two extremes. Labor supply often is elastic, but not perfectly elastic. Fertilizers and pesticides are other examples, as these inputs are subsidized up to a fixed quotum. Two types of factors are distinguished in the model: those supplied at the level of the farm-type land unit, i.e., land, water from sources other than the river, family labor, farm-owned machinery, and subsidized fertilizers and pesticides, and those supplied at the sub-regional level, i.e., hired labor, river water, services from rented agricultural machinery, and non-subsidized fertilizers and pesticides.

All purchased inputs and services are priced at observed prices, except those that are explicitly subsidized, such as fertilizers and pesticides. For the latter, both subsidized and non-subsidized (or market) prices are included. For resource inputs whose availability is fixed in the short run, such as land, water, and family labor, the question arises whether they should be priced explicitly in the primal version of the model (Hazell and Norton, 1986). For land, the implicit opportunity cost is represented by its productivity in the most remunerative agricultural use. Agricultural land is not priced, as it is assumed that it has no value outside agriculture in the short run, but the solution of the model yields the value that accrues to the land. Similarly, the water is not priced, but the costs of tapping the water supply and providing it to farms are included in the production costs that are charged against the objective function.

Hired labor wages are set to the current market levels for each of the farm-type land units. For inputs of family labor that are not explicitly paid for, a minimum expected return is assumed, which is often referred to as the "reservation wage," and is difficult to assess *a priori*. That return, or the implicit wage, almost certainly exceeds zero, but is also likely to be below the market wage. Analyses by Duloy and Norton (1990) and Bassoco and Norton (1990) suggest that, for Mexico, the reservation wage for family labor is 30–70% of the market wage. The current model has been structured in such a way that the ratio of farmers' reservation wage to hired labor can be introduced exogenously.

The manageability of sector models is enhanced considerably, if output pricing and input costing activities are kept separate from land-use activities, even though this requires use of additional input and output balances (Hazell and Norton, 1986). Therefore, in this chapter, frequent use has been made of separate balances. The advantages of this specification are: input supplies can be both costed and bounded if appropriate; multi-step, upward sloping input supply functions can be introduced; and changes in input prices or supply conditions can be relatively easily introduced, often by changing one parameter in the model, instead of hundreds or thousands of aggregate cost coefficients for all production vectors. An additional advantage is the transparency of the structure in the tableau.

To allow analysis of the impact of the policy environment in the model, various policy objectives have been included. Various land-use scenarios corresponding to different policy measures are defined. Policy measures (policy instruments) are represented in the model structure by a set of parameters: coefficients in the matrix, the right hand side, and/or the objective function. The policy instruments have been

Table 13.2. Policy objectives used in the evaluation procedure.

Realm of objective	Objective
Agro-technical	Increase in overall production efficiency
	Increase in food production efficiency
Socio-economic	Increase in total farm income
	Decrease in total variable production cost
	Attain equitable income distribution
	Increase total "steady" employment
	Decrease total "highly-seasonal" employment
Agro-ecological	Decrease nitrogen losses per hectare
	Decrease input of pesticide per hectare

tested by solving the model under alternative assumptions with respect to the values of the policy parameters. On the basis of these scenarios, the model simulates the impact of policy changes on the various policy objectives.

13.3.4 Generation of land-use policy scenarios

To represent the multiple and (partially) conflicting views of different stakeholders, various agro-technical, socioeconomic, and agro-ecological objectives have been identified, of which nine are finally used in the evaluation procedure (*Table 13.2*). The grouping of these objectives is not unequivocal; for instance, some of the objectives included in the socioeconomic realm may have an agro-technical dimension or vice versa. However, this grouping has no effect on the evaluation procedure.

To achieve policy objectives, instruments need to be identified that influence farmers' decisions on land use and allocation of other resources. Five policy instruments are considered: increase in price of rice (improved variety), fertilizer subsidy withdrawal, pesticide subsidy withdrawal, mechanization of rice transplanting and harvesting activities, and consolidating the land. Each of these policy measures is expected to contribute to achievement of some of the policy objectives. They also represent perceptions of different stakeholders with respect to the policy interventions. Policy instruments have been identified on the basis of discussions with resource persons in the Amol region, including regional policy makers, farmers' representatives, co-operative societies, and other regional organizations.

Policy measures or instruments are represented in the model structure by coefficients in the matrix, the right hand side, and/or the objective function. Following generation of a base solution, various land-use policy scenarios are defined, corresponding to various policy instruments, and have been incorporated in the model through introduction of new values for the parameters, characterizing these policy instruments. By solving the model with these new values of the variables, a set of land-use policy scenarios is generated, each showing the relation between a policy instrument and its effects on policy objectives.

Various policy instruments are translated into different policy scenarios. In this chapter, the linear programming model has been run for six scenarios reflecting various policy instruments:

1. the base scenario (BSCN),
2. price of improved rice scenario (SCN1),
3. fertilizer subsidy withdrawal scenario (SCN2),
4. pesticide subsidy withdrawal scenario (SCN3),
5. mechanization of rice transplanting and harvesting scenario (SCN4),
6. land consolidation scenario (SCN5).

The results are introduced in the policy impact matrix, a two-dimensional matrix including policy instruments (alternatives) and policy objectives (criteria). Each entry (effect score) in the matrix represents the consequence of a specific alternative for each criterion. The model results point to large differences in the values of the policy objectives among the six policy scenarios. The results show that there is indeed no one single policy scenario that results in the 'best' values for all policy objectives, justifying a need for a multi-criteria evaluation.

13.3.5 Evaluation of land-use policy scenarios

Before carrying out a multi-criteria evaluation, the criteria scores must be standardized to enable meaningful comparisons on the basis of criteria expressed in different units, for which various procedures exist (see for example Voogd, 1983). Another important step in most multi-criteria evaluation methods is assignment of weights or priorities, reflecting the (relative) importance attached to the various impacts by the user, or, in more general terms, assessment of a preference structure. In formulating and assessing preferences, the limitations in human capabilities in this respect should be considered. It is not realistic to expect policy makers to be able to quantify the policy preferences (weights) among objectives, in advance. They are often not prepared or unable to formulate their priorities explicitly. Moreover, in scenario studies, it may be inappropriate to start multi-criteria evaluation with a unique representation of policy priorities.

A useful alternative is to identify a number of different combinations of priorities, each of which is representative for a possible policy view (Veeneklaas *et al*., 1991). In this chapter, therefore, hypothetical, qualitative priority statements – linked to a particular policy view – are used. These priorities are represented through ordinal expressions (e.g., more important, equally important, less important, etc.), rather than by weights, which are represented by numerical expressions

Table 13.3. Policy views used for the evaluation of policy scenarios.

Policy view	Description
Welfare of farmers	High priority for increasing farm income and production efficiency, low costs, and reduction in high seasonality in labor demand
Regional development	High priority for increasing regional employment, more equitable income distribution, and food production efficiency
Environmental protection	High priority for the reduction in contaminants from the agricultural sector to the environment
Compromise	Equal priority for all policy objectives considered in the evaluation process

(e.g., 0.15, 0.20, 0.65). By showing the consequences of various policy views, the model can provide an objective basis for debate on these preferences (Hazell and Norton, 1986), thus assisting policy makers in selecting the most preferred alternative or to facilitate a movement towards a consensus.

In the context of regional development, four possible views on priorities with respect to policy objectives have been formulated, as summarized in *Table 13.3*. These policy views are widely divergent, to arrive at contesting policy visions or aspirations and to represent the major opinions on the desired development of the region. Each of the views prioritizes some of the objectives, while conceding that other objectives are also valid. In environmental protection, for instance, attention is focused on reducing emission of harmful substances from agriculture to the environment, which requires minimization of the use of fertilizers and pesticides.

Many different multi-criteria evaluation methods or techniques exist that can be used for evaluating alternatives (Voogd, 1983; Janssen, 1992). The general advice is to use a small number of alternatives and only a limited number of criteria, ideally of the order of eight for each (Voogd, 1983). In this chapter, various methods have been applied for ranking the alternative policy scenarios under the four specified policy views. The purpose of this appraisal is to answer the question: which policy instrument(s) is/are suitable (or desirable) for achieving specific policy objective(s) under each of the specified policy priorities? The answer should provide the policy makers with a menu of policy instruments, with their consequences for policy objectives, under different assumptions with respect to desired policy directions and priorities. The rankings of the various policy scenarios, from different policy perspectives, produced by DEFINITE software, are presented in *Table 13.4*.

The results show that for the specific situation of the Amol region, and under the assumed policy views: (a) non-price policy instruments are more effective in bringing about desired changes and in achieving policy objectives; and (b) when priority is given to environmental protection, the current situation, as reflected in

Table 13.4. Summary of the ranking of the alternative policy instruments from different policy views.

		Policy scenario[c]					
Policy view[a]	Evaluation method[b]	BSCN	SCN1	SCN2	SCN3	SCN4	SCN5
Welfare of farmers	Weighted summation	3	4	5	6	1	2
	Regime	4	3	5	6	1	2
	Expected value	3	4	5	6	1	2
	Evamix	3	4	5	6	1	2
Regional development	Weighted summation	4	1	5	6	3	2
	Regime	4	2	5	6	1	3
	Expected value	4	1	5	6	3	2
	Evamix	4	1	5	6	3	1
Environmental protection	Weighted summation	6	5	4	2	1	3
	Regime	6	4	5	2	1	3
	Expected value	6	5	4	2	1	3
	Evamix	6	5	4	2	1	3
Compromise	Weighted summation	4	3	5	6	2	1
	Regime	4	3	5	6	2	1
	Expected value	4	3	5	6	2	1
	Evamix	4	3	5	6	2	1

[a] Policy view: description of policy views is given in *Table 13.3*.
[b] Evaluation method: Weighted summation method, regime method, expected value method, and Evamix method, are evaluation methods that have different arithmetic procedures for combining the information from the evaluation matrix with the information contained in the priority matrix. This results in an appraisal matrix that gives an indication of the ranking of alternatives. For more details on these methods the reader is referred to Voogd (1983) or Janssen (1992).
[c] For description of the policy scenarios, see text.

the base scenario (BSCN), is ranked most unfavorably; the 'land consolidation' scenario (SCN5) is a good compromise among the different policy views.

13.4 Assessment of the Approach

The usefulness of the model developed in this chapter, and the relevance of its results have been evaluated on the basis of its performance to a real case study. To assess the quality of the model, the generated results, in terms of the magnitude of the aggregation errors, have been compared to those obtained from other land-use modeling approaches. Aggregation errors have been quantified by comparing observed data to the results of three approaches of modeling land use in Amol Township, each with a different spatial unit as basis for the analysis. The three approaches include:

Table 13.5. Quantification of aggregation errors using different modeling approaches.

	Actual	Estimates by the various modeling approaches			% Aggregation errors in the various modeling approaches		
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fertilizer input use							
urea (tons)	11,334	11,886	13,424	14,764	4.9	18.4	30.3
diammoniumphosphate (tons)	7,950	8,125	9,386	9,913	2.2	18.1	24.7
Production							
tarom (tons)	79,695	77,068	71,814	56,724	–3.3	– 9.9	–28.8
amol3 (tons)	73,316	77,833	97,039	127,660	6.2	32.4	74.1
Land use							
tarom (ha)	20,700	19,999	15,791	14,102	–3.4	–23.7	–31.9
amol3 (ha)	12,490	12,829	16,953	19,559	2.7	35.7	56.6

1. a sub-regional model based on farm-type land units (Model 1),
2. a sub-regional model based on farm types (Model 2),
3. a sub-regional model in which the sub-region is modeled as a single farm (Model 3).

In terms of production techniques, only current land-use systems were included. In all cases, the objective function is maximization of net benefits under 1994 input and output prices.

In general, performance of a model is judged by comparison of its results to those of the model based on the most disaggregated units (see, for example, Jansen and Stoorvogel, 1998). This procedure is based on the implicit (biased) assumption, that the lower the level of aggregation, the smaller the error. However, that is not necessarily true. In this chapter, the available data on the current situation are sufficiently detailed to be used as a yardstick against which the estimates obtained from the three models can be judged. Estimates obtained from three sub-regional models, corresponding to the different approaches, are thus compared with the actual situation. The percentage deviation of the estimates of the models from the actual situation is defined as aggregation error. This aggregation error is quantified for each of the three approaches in terms of land use, input use, and crop production, as presented in *Table 13.5*.

Model 1 shows the lowest aggregation errors, Model 3 the highest. The aggregation errors in Model 1 are indeed small compared to those in both Model 2 and Model 3. The aggregation error in Model 3 is considerable, with, for instance, an estimated production of improved rice (amol3) exceeding actual production by more than 74%. Aggregation errors in Model 2 are, on average, 58% of those in Model 3, while those in Model 1 are, on average, 20% and 10% of those in Model 2 and Model 3, respectively.

13.5 Discussion and Conclusions

The proposed modeling approach requires an interdisciplinary approach to land-use planning and policy analysis, because it integrates information from socioeconomic and agro-ecological disciplines, and combines various tools and techniques, derived from many specializations. It allows generation and evaluation of policy scenarios, associated with different policy instruments and their consequences for development policy objectives. It allows evaluation of these scenarios under different policy priorities.

A major advantage of this approach is, that it can address the issue of effects of spatial scales. It includes three spatial aggregation levels: farm-type land unit level, farm-type level, and sub-regional level. The farm-type land unit has been selected as the unit for land-use modeling. This use of FTLU in land-use planning and policy analysis facilitates the integration procedure by combining socioeconomic and biophysical aspects in the description and quantification of land-use systems. It simulates the behavior of farmers at the micro level; it permits analysis of the impact of policy options at various levels – FTLU, FT, and sub-region – thus relating macro and micro levels; and it reduces aggregation bias by allowing restrictions on resource mobility, by allowing for differences in technologies of production, and by incorporating spatial differentials in prices and costs.

Data collected and databases created during farming systems research are generally insufficiently used for policy simulations at local and regional levels (Van Keulen *et al*., 1998). An important characteristic of the present approach is the design of a large farming systems database, designed to easily retrieve data in all required sequences and formats, and to perform tasks of data analysis and presentation without losing detail stored in the information. Moreover, the inclusion of location parameters as attributes in the farming systems database permits explicit geo-referencing and subsequent linking of socioeconomic data with the biophysical data, using a combination of GIS and statistical methods. Integration of socioeconomic information, obtained from farming systems analysis and biophysical information derived from land evaluation, has often met with great difficulties, because of the lack of geo-referencing of the former. Hence, the proposed explicit geo-referencing (or mapping) of the farm types, is indeed a significant improvement in land-use planning approaches, as it provides an essential missing link between the socioeconomic and agro-ecological information.

The approach forms the basis for improved interaction between agricultural research and information management. By identifying data needs and requirements for effective agricultural planning and policy analysis, the methodology may guide data collection and stimulate development of improved appropriate databases and information systems. As a consequence, it directs agricultural research by identifying information gaps. A strong point of the methodology is the use of a quantitative approach in the integration of agro-ecological and socioeconomic information, including explicit formulation of the assumptions and relationships underlying these quantifications, in support of policy analysis. Hence, the methodology is more transparent than many qualitative tools. However, various aspects of land-use decisions are less easy (if not impossible) to quantify. This applies to many sociological, cultural, and even ecological variables and relationships, which can play an important (or even decisive) role.

Further development of the approach is required, however, to deal with major limitations, such as absence of temporal variation, absence of the risk dimension in farm household decision making, absence of the consumption side in farm household modeling, absence of factor substitution possibilities in the quantification of land-use systems, and absence of the increasingly recognized role of non-agricultural income in farm household decision making. Moreover, a major limitation of the proposed methodology is also recognized: its operationalization requires a very large amount of data at farm level. These data are, in general, not available in many developing countries, which moreover may lack the necessary resources to collect them. Nevertheless, the approach can be implemented with less data, and can be used to indicate information gaps that can be filled gradually.

In conclusion, the proposed approach considerably reduces the aggregation errors, when compared to the existing modeling approaches in land-use planning and policy analysis and is therefore expected to make a significantly positive contribution to improved quality of agricultural planning and policy analysis. Some degree of aggregation is, of course, inevitable in modeling and necessary to restrict the costs of the analysis to 'reasonable' levels. Implementation of the proposed methodology requires a large database and the gains in precision of the analysis must be balanced against the higher costs of developing and implementing the methodology.

References

Alfaro, R., Bouma, J., Fresco, L.O., Janssen, D.M., Kroonenberg, S.B., Leeuwen, A.C.J. van, Schipper, R.A., Sevenhuysen, R.J., Stoorvogel, J.J., and Watson, V., 1994, Sustainable land use planning in Costa Rica: a methodological case study on farm and regional level, in L.O. Fresco, L. Stroosnijder, J. Bouma, and H. van Keulen, eds, *The Future of the Land: Mobilizing and Integrating Knowledge for Land Use Options*, John Wiley and Sons, Chichester, UK, pp. 183–202.

Bassoco, L.M., and Norton, R.D., 1990, A quantitative approach to agricultural policy planning, in E. Thorbecke and L. Hall, eds, *Agricultural Sector Analysis and Models in Developing Countries*, FAO Economic and Social Development Paper 5, FAO, Rome, Italy, pp. 263–286.

Beek, K.J., 1978, Land evaluation for agricultural development: some exploration of land use system analysis with particular reference to Latin America, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, Netherlands.

Benyon, D., 1990, *Information and Data Modelling. Information Systems Series,* Blackwell Scientific Publications, Oxford, UK.

Braat, L.C., and Lierop, W.H.J. van, 1987, Integrated economic-ecological modelling, in L.C. Braat and W.H.J. van Lierop, eds, *Economic-Ecological Modelling. Studies in Regional Science and Urban Economics*, Volume 16, North Holland, Amsterdam, Netherlands, pp. 49–68.

Conway, G.R., 1987, The properties of agroecosystems, *Agricultural Systems*, **24**:95–117.

Date, C.J., 1990, *An Introduction to Database Systems*, Volume 1, 4th edition, Addison Wesley, New York, NY, USA.

Day, R.H., 1963, On aggregating linear programming models of production, *Journal of Farm Economics,* **45**:797–813.

Diepen, C.A. van, Keulen, H. van, Wolf, J. and Berkhout, J.A.A., 1991, Land evaluation: From intuition to quantification, *Advances in Soil Sciences*, **15**:139–204.

Driessen, P.M., and Konijn, N.T., 1992, Land Use Systems Analysis, Wageningen Agricultural University, Department of Soil and Geology, Malang: INRES.

Duivenbooden, N. van., Gosseye, P.A., and Keulen, H. van, 1991, Competing for limited resources: the case of the fifth region of Mali, Report 2, Plant, livestock and fish production, CABO/ESPR, Wageningen/Mopti.

Duloy, J.H., and Norton, R.D., 1990, CHAC, a programming model of Mexican agriculture, in E. Thorbecke and L. Hall, eds, *Agricultural Sector Analysis and Models in Developing Countries*, FAO Economic and Social Development Paper 5, FAO, Rome, Italy, pp. 227–262.

Erenstein, O.C.A, and Schipper, R.A., 1993, Linear programming and land use planning: A case study of Matara district, Sri Lanka, Wageningen Economic Studies 28, Department of Development Economics, Wageningen Agricultural University, Wageningen, Netherlands.

FAO, 1976, A framework for land evaluation, FAO Soils Bulletin 32, FAO, Rome, Italy.

FAO, 1983, Guidelines. Land evaluation for rainfed agriculture, FAO Soils Bulletin 52, FAO, Rome, Italy.

FAO, 1984, Land evaluation for forestry, FAO Forestry Paper 48, FAO, Rome, Italy.

FAO, 1985, Guidelines. Land evaluation for irrigated agriculture, FAO Soils Bulletin 55, FAO, Rome, Italy.

FAO, 1987, Guidelines. Land evaluation for extensive grazing, Final Draft, FAO, Rome, Italy.

Fresco, L.O., 1986, Cassava in shifting cultivation. A system approach to agricultural technology development in Africa, Royal Tropical Institute, Amsterdam, Netherlands.

Fresco, L., 1994, Planning for the people and the land of the future, in L.O. Fresco, L. Stroosnijder, J. Bouma, and H. van Keulen, eds, *The Future of the Land: Mobilizing and Integrating Knowledge for Land Use Options*, John Wiley and Sons, New York, NY, USA, pp. 395–398.

Fresco, L.O., Huizing, H.G.J., van Keulen, H., Luning, H.A., and Schipper, R.A., 1992, Land evaluation and farming systems analysis for land use planning, FAO working document.

Hazell, P.B.P., and Norton, R.D., 1986, *Mathematical Programming for Economic Analysis in Agriculture*, Macmillan, New York, NY, USA.

Heemst, H.D.J. van, 1986, Crop calendar, workability and labour requirements, in H. van Keulen and J. Wolf, eds, *Modelling of Agricultural Production: Weather, Soils and Crops*. Pudoc, Wageningen, pp. 251–262.

Heemst, H.D.J. van, Merkelijn, J.J., and Keulen, H. van, 1981, Labour requirements in various agricultural systems, *Quarterly Journal of International Agriculture*, **20**(2):178–201.

Hengsdijk, H., and Kruseman, G., 1993, Operationalising the DLV program: an integrated agro-economic and agro-ecological approach to a methodology for analysis of sustainable land use and regional agricultural policy, DLV Report No. 1, AB-DLO/Department of Development Economics, Wageningen Agricultural University, Wageningen, Netherlands.

Howe, D.R., 1989, *Data Analysis for Data Base Design*, Edward Arnold, London, UK.

Huizing, H., and Bronsveld, K., 1994, Interactive multiple-goal analysis for land use planning, *ITC Journal*, **4**:366–373.

Ittersum, M.K. van, 1998, New concepts and directions in exploratory land use studies, in R. Rötter, C.T. Hoanh, N.V. Luat, M.K. van Ittersum and H.H. van Laaar, eds, *Exchange of Methodologies in Land Use Planning*, SysNet Research Paper Series No. 1, IRRI, Los Banos, Philippines, pp. 3–13.

Ittersum, M.K. van, and Rabbinge, R., 1997, Concepts in production ecology for the analysis and quantification of agricultural input-output combinations, *Field Crops Research*, **52**:197–208.

Janssen, R., 1992, *Multiobjective Decision Support for Environmental Management, Environment and Management*, Volume 2, Kluwer Academic Publishers, Dordrecht, Netherlands.

Jansen, D.M., and Schipper, R.A., 1995, A static, descriptive approach to quantify land use systems, *Netherlands Journal of Agricultural Sciences*, **43**:31–46.

Jansen, H.G.P., and Stoorvogel, J.J., 1998, Quantification of aggregation bias in regional agricultural land use models: Application to Guàcimo County, Costa Rica, *Agricultural Systems*, **58**:417–439.

Keulen, H. van, Kuyvenhoven, A., and Ruben, R., 1998, Sustainable land use and food security in developing countries: DLV's approach to policy support, *Agricultural Systems*, **58**:285–307.

King, P., 1995, Soil survey and land evaluation consultancy report, in *Land Use Planning II*, Islamic Republic of Iran, Western Australian Department of Agriculture, Perth, WA, Australia.

Koning, G.H.J. de, Janssen, H., and Keulen, H. van, 1992, Input and output coefficients of various cropping and livestock systems in the European Community, Working document 62, Netherlands Scientific Council for Government Policy, The Hague, Netherlands.

Lanen, H.A.J. van, 1991, Qualitative and quantitative physical land evaluation: an operational approach, Ph.D. thesis, Wageningen Agricultural University, Wageningen, Netherlands.

Latesteijn, H.C. van, 1999, Land use in Europe. A methodology for policy-oriented future studies, Preliminary and Background Studies V106, WRR, Sdu Uitgevers, The Hague, Netherlands.

Luning, H.A., 1986, Survey integration comes of age? International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede, Netherlands.

Malingreau, J.P., and Mangunsukardjo, K., 1978, Land evaluation and integrated approach to rural development, Interpretasi Citra Penginderaan Jauh dan Survey Terpadu. UGM, Bakosurtanal.

Mohamed, A.A., 2000, An integrated agro-economic and agro-ecological framework for land use planning and policy analysis, Ph.D. thesis, Wageningen University, Wageningen, Netherlands.

Mohamed, A.A., and Sharifi, M.A., 1998, In search for an integrated interdisciplinary unit for land use planning and policy analysis: conceptualization and operationalization, in L.H. Kevin and K. Demetri, eds, *Our Natural Environment: Interdisciplinary Interactions. Proceedings of the Fourth International Interdisciplinary Conference on the Environment*, Washington D.C., USA, July 1–10, 1998, pp. 164–177.

Pichett, S.T.A., Burke, I.C., Dale, V.H., Gosz, J.R., Lee, R.G., Pacala, S.W., and Schachak, M., 1994, Integrated models of forested regions, in P.M. Groffman and G.E.Likens, eds, *Integrated Regional Models: Interactions Between Humans and Their Environment*, Chapman and Hall, London, UK, pp. 120–141.

Polman, F.J., Samad, M., and Thio, K.S., 1982, A demonstration of resource based socio-economic planning in Matara district, Report No 1. in Regional Planning for Agricultural Development in Sri Lanka, Agrarian Research and Training Institute, Colombo, Sri Lanka.

RAWOO, 1989, Sustainable land use in developing countries. Perspectives on an integrated approach, Working Paper 2, RAWOO, The Hague, Netherlands.

Rheenen, T. van, 1995, Farm household level optimal resource allocation. An explorative study in the limestone area of East Java, Ph.D. thesis, Wageningen Agricultural University, Wageningen, Netherlands.

Schipper, R.A., 1996, Farming in a fragile future. Economics of land use with applications in the Atlantic Zone of Costa Rica, Ph.D. thesis, Wageningen Agricultural University, Wageningen, Netherlands.

Schipper, R.A., Janssen, D.M., and Stoorvogel, J.J., 1995, Sub-regional linear programming models in land use analysis: A case study of the Neguev settlement, Costa Rica, *Netherlands Journal of Agricultural Sciences*, **43**:83–109.

Sharifi, M.A., and Keulen, H. van, 1994, A decision support system for land use planning at farm level, *Agricultural Systems*, **45**:239–257.

Stomph, T.J., and Fresco, L.O., 1991, Describing agricultural land use. A proposal for procedures, a data base and a users' manual to be incorporated in a FAO Soils Bulletin, draft, FAO, Rome, Italy.

Stomph, T.J., Fresco, L.O., and Keulen, H. van, 1994, Land use system evaluation: concepts and methodology, *Agricultural Systems*, **44**:243–255.

Stoorvogel, J.J., 1995, Linking GIS and models: Structure and operationalisation for a Costa Rican case study, *Netherlands Journal of Agricultural Sciences*, **43**:19–29.

Taazimi, M., 1995, Climatic resource inventory: Mazandaran region, in *Land Use Planning II*, MOA/UNDP/FAO, Center for Agricultural Planning and Economic Studies, Ministry of Agriculture, Islamic Republic of Iran, Tehran, Iran.

Thorbecke, E., and Hall, L., 1990, Nature and scope of agricultural sector analysis: an overview, in E. Thorbecke and L. Hall, eds, *Agricultural Sector Analysis and Models in Developing Countries*, FAO Economic and Social Development Paper 5, FAO, Rome, Italy, pp. 3–13.

Van Dyne, G.M., and Abramsky, Z., 1975, Agricultural systems models and modeling: an overview, in G.E. Dalton, ed., *Study of Agricultural Systems*, Applied Science Publishers, London, UK, pp. 23–106.

Veeneklaas, F.R., Cissé, S., Gosseye, P.A., and Keulen, H. van, 1991, Competing for limited resources: The case of the fifth region of Mali, Report 4: Development Scenarios, CABO-DLO/ ESPR.

Voogd, H., 1983, *Multicriteria Evaluation for Urban and Regional Planning*, Pion Ltd, London, UK.

Zonneveld, I.S., 1997, *Land Ecology: An Introduction to Landscape Ecology as a Base for Land Evaluation, Land Management and Conservation*, SPB Academic Publishing, Amsterdam, Netherlands.

Chapter 14

The Importance of Cattle Grazing in Land Resource Management: The Case of the Mount Sanbe Grassland in Southwestern Japan

Yoshitaka Takahashi

Abstract

Mount Sanbe in southwestern Japan was once characterized by a pastoral landscape which extended from the foot of the mountain to its summit, and had been sustained for 400 years through grazing management. Recognized for its natural beauty, Mount Sanbe was designated part of the Daisen-Oki National Park in 1963. However, the replacement of cattle by farming machines coupled with an increase in the number of cattle raised indoors, resulted in decreased grazing activity on Mount Sanbe, and the grasslands began to be neglected. In 1996, after a quarter century of interruption, cattle grazing was revived on the west foot (Nishinohara) of the mountain. The revival of cattle grazing has not only promoted local livestock farming and rural tourism, it has also promoted the ecological restoration of this pastoral landscape, conservation of its biodiversity, and the creation of a space where people and animals can commune together in nature. Herein, several cases are introduced to demonstrate the multiple effects of livestock grazing, and a discussion follows regarding the significance of grazing in rural development.

Keywords: Cattle grazing, conservation, land resource, semi-natural grassland, utilization.

14.1 Introduction

Grassland is a particularly good example of land use in which agriculture can co-exist with highly natural and scenic areas (Green, 1996; Numata, 1994). Until relatively recently, Mount Sanbe in the Shimane Prefecture in Japan was a region in which people were a closely integrated part of the sustainable agricultural system.

Mount Sanbe was once characterized by a pastoral landscape running from the foot of the mountain to its summit, which had been sustained under grazing management for 400 years (Shoji *et al*., 1995; Takahashi, 1994). Recognized for its natural beauty, it was designated as a part of the Daisen-Oki National Park in 1963. However, the replacement of cattle by machines for farming and the increase in the number of cattle raised indoors resulted in decreased grazing activity on Mount Sanbe, and the grasslands were no longer managed or considered to be important.

After a quarter century of neglect, however, cattle grazing was revived in 1996 at the west foot (Nishinohara) of the mountain (Takahashi, 1997). The revival of cattle grazing has not only promoted local livestock farming and rural tourism, but ecological restoration of a pastoral landscape, conservation of biodiversity, and the creation of amenity space with landscape animals. This example may provide a useful case study on the necessity for the integration of agricultural, regional, and environmental aspects for land resource management.

In this chapter, several cases are introduced to demonstrate the multiple effects of livestock grazing on Mount Sanbe. A discussion regarding the significance of grazing to environment management and rural development follows.

14.2 Semi-Natural Grasslands in Japan

Japan is located in a heavy rainfall zone, and its climate is warm and humid. Therefore, it contains no natural grasslands where grasses exist as the climax vegetation except for alpine grasslands, windward grasslands, coastal grasslands, and moors that cover only very limited areas (*Table 14.1*). Most of the grasslands in Japan are regarded as secondary grasslands (semi-natural grasslands), which are maintained under grazing, mowing, and burning management (Numata, 1994; Takahashi and Naito, 1997).

Table 14.1. Criteria of the degree of vegetation naturalness and its proportion in Japanese natural parks (%, 1989).

Degree of naturalness	Vegetation and landscape	National park	Quasi-national park	Japan
10	Natural grassland	4.5	2.7	1.1
9	Climax forest	51.3	31.9	18.2
8–7	Secondary forest	16.1	23.1	24.5
6	Plantation	14.8	23.8	24.7
5–4	Secondary grassland	4.2	3.8	3.2
3–2	Cultivated field	4.4	4.6	22.7
1	Urban area, preparation site	0.8	0.8	4.0
Others		4.0	9.3	1.6

Source: The Environment Agency, 1989.

According to the survey of the Environment Agency (*Table 14.1*), managed grasslands in Japan cover only 3% of the total land area. A hundred years ago, grasslands covered more than 11% of the land area in Japan. In order to use grasses to fertilize the paddy fields, it was necessary to have grassland areas that were larger than the paddy fields.

Some grasslands remain today in national and quasi-national parks. However, as can be seen in *Table 14.1*, grasslands cover only a very small percentage of the total land area of Japan. To preserve the cultural landscape and possibly to maintain biodiversity, we believe that 20% of the nation's grasslands should be included within park boundaries. It is therefore of importance to maintain the grazing pastures and meadows at and around Mount Sanbe, which was designated as a part of Daisen-Oki National Park in 1963.

14.3 Changes in Grassland Vegetation and Cattle Grazing in Mount Sanbe

Mount Sanbe is a volcano located in the central Shimane Prefecture in Southwestern Japan, the top of which is 1,126 meters (m) above sea level. The average temperature is 13°C at the foot of the mountain (400–500 m above sea level). Mount Sanbe hosts a ski resort and is located near the Sea of Japan.

In the past, the mountain was used as a pasture, and more than a thousand cows were pastured there (*Photo 14.1*). In those days, each farm had many cows in order to utilize the down slope for livestock husbandry and organic fertilizer. However, after World War II, the role of cattle shifted from draft purposes to beef production. Raising cattle in sheds is now more common than grazing them in pastures.

Photo 14.1. View of the west pasture in the 1930s.

Photo 14.2. View of the west pasture in the 1960s.

Until about the 1960s, there were many grasslands in the Mount Sanbe area (*Photo 14.2*). One of the reasons for the Mount Sanbe area to be included in the Daisen-Oki National Park was its beautiful scenery, which included cows grazing on the grasslands (Shoji *et al*., 1995; Takahashi, 1994).

Unfortunately, the number of grazing cows has decreased, while forestation has increased; in the process, several attractive flowering plants have also been lost. Recently, most of the grasslands have been left unused, and the areas have been replaced by forest (*Photo 14.3*). As a whole, we have 600 hectares (ha) of

Photo 14.3. View of the abandoned pasture in 1995.

grasslands, most of which have resulted from forest clear-cut stands and abandoned fields (*Figure 14.1*). Only 200 ha of grasslands are presently used for livestock farming.

14.4 The Revival of Cattle Grazing and its Significance

In recent years, there has been a movement towards changing production systems that had begun to rely upon intensively raised livestock. The public often thinks that the health and welfare of extensively raised animals is better than that of their intensively raised counterparts. This is because the public places a high value on animals being allowed to live under natural conditions and having the freedom to present normal behavior.

In the Mount Sanbe area, extensive cattle grazing was re-evaluated as a way to utilize grass resources in broad pastures. It was thought that a return to grazing practices could result in labor savings and reduced costs in the production of beef cattle. In addition, the problem of losing the special landscape of Mount Sanbe, characterized by cattle grazing and pastures, had begun to be recognized as a serious problem; therefore it began to seem desirable to revive grazing and thus conserve the grassland landscape (Takahashi, 1997).

After a quarter century of obsolescence, cattle grazing was revived in 1996 at the west foot (Nishinohara) of the mountain. This change was originally proposed

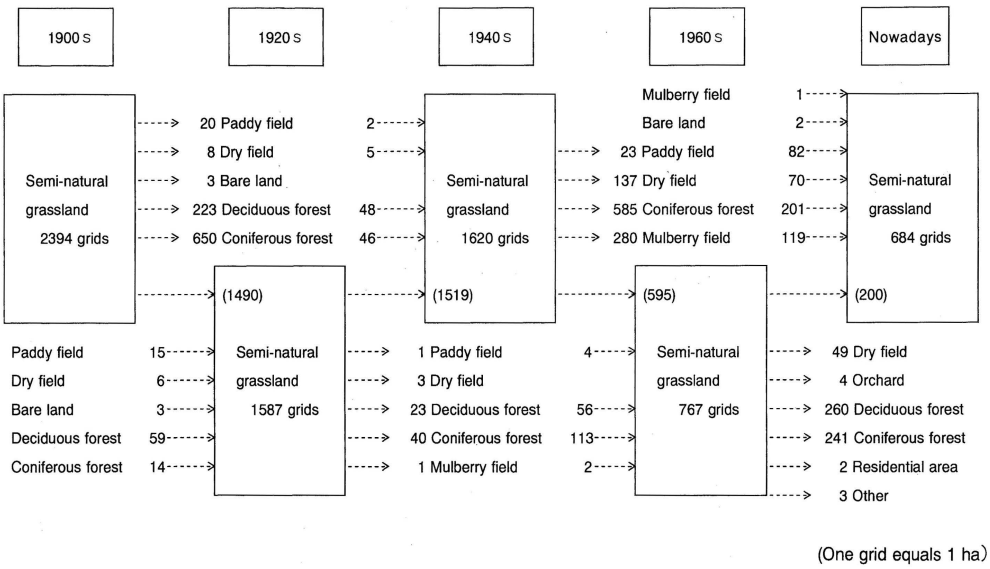


Figure 14.1. Changes in land use in the Mount Sanbe area (Shoji *et al.*, 1995).

by local farmers and scholars for the revival of local farming, and was later supported by the local government as a tourist attraction and by citizen's organizations concerned with nature conservation.

The revival of cattle grazing has promoted not only local livestock farming and rural tourism, but also ecological restoration of the pastoral landscape, the conservation of biodiversity, and the creation of amenity space with landscape animals.

14.4.1 Promotion of local livestock farming

In Nishinohara, the west pasture, grazing was revived in 1996, and the number of cattle has increased since then. In the first year, the grazing area was 24 ha, and in the second year an additional area of 20 ha was added, making the final pasture size approximately 45 ha. The total number of cattle in the Mount Sanbe area has increased as that in Nishinohara has increased (*Figure 14.2*). The carrying capacity of the grasslands also increased as the vegetation changed from a *Miscanthus*-type to a *Zoysia*-type grasslands, a change that makes it more effective for grazing.

The farmers pay only 2,000 yen per head per year to the Sanbe Pasture Committee for their cattle grazing, which is very little. High costs are not necessary to maintaining a Zoysia grassland; it can be maintained by cattle grazing. In addition, it offers the tourists a very attractive uniform surface. Neighboring farmers

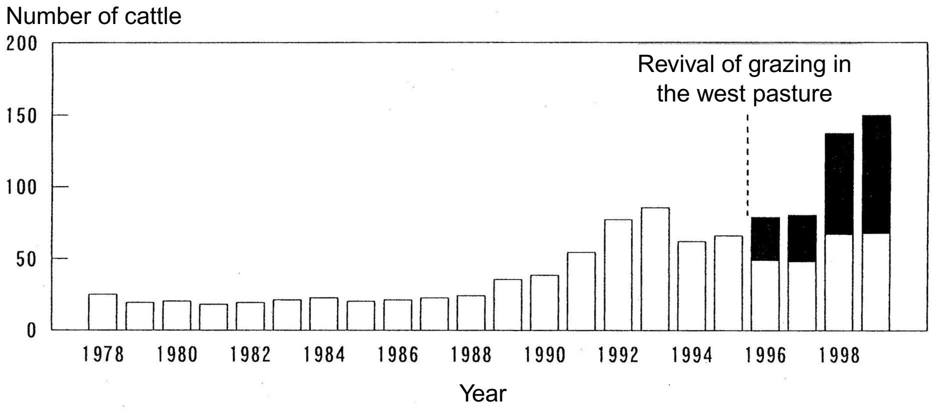


Figure 14.2. Changes in the number of cattle in pastures in the Mount Sanbe area (a closed column indicates the number in the west pasture).

are interested in adopting a grazing system and enlarging the scale on which they are raising beef cattle.

The grazing period generally lasts from April to almost the end of December. During the grazing season, farmers are able to prepare forages for winter or do other jobs such as rice and vegetable cultivation, so they can actually do all of the above work throughout the year. Their management thus becomes more highly developed and more profitable.

14.4.2 Ecological restoration of the pastoral landscape

The cessation of cattle grazing meant that the grasslands were taken over by shrubs and saplings covering the *Miscanthus sinensis* grass layers (*Photo 14.3*). Thanks to the "mowing" carried out by the cattle, however, the beautiful grassland landscape has been slowly but steadily restored since the revival of grazing.

After just 3 years since the beginning of the revival, the traditional landscape of the *Zoysia*-type grassland on Mount Sanbe, which was one of its essential characteristics for being designated as a national park, had begun to reappear. The vigorous appetite of cattle has inhibited the tall grasses and their dung has spread the *Zoysia* seeds, whose germination is improved by passing through the stomach of cattle. Thus the pasture had changed to short turf, which looks like a green carpet. During the course of this change, some grassland species, which could not have survived under the cover of tall plants, have also increased their populations.

Rhododendron japonicum is the city flower of Oda, where this grassland is located, and the citizens are quite proud of this flower. When grazing stopped and the grasslands were disappearing, *R. japonicum* was infrequently seen due to its

Photo 14.4. *Rhododendron japonicum* in the west pasture.

inhibition by tall plants. The revival of grazing, however, which has resulted in the transformation to a *Zoysia japonica* grassland, has also increased the population of *R. japonicum*. We can now see more *R. japonicum* in the grasslands because cattle eat the tall grasses that surround each *R. japonicum* plant (*Photo 14.4*).

Before grazing was started in 1996, field litters caught fire frequently, especially in winter and early spring. In order to prevent mountain fires, it was decided to carry out annual burnings. However, in the pastures, there has been so little litter that it is unnecessary to conduct burnings.

14.4.3 Ecological lawn mower for biodiversity maintenance

The disappearance of traditional grazing practices endangered biodiversity in the grasslands of Mount Sanbe (Takahashi and Naito, 1997). When grazing was abandoned, the plants became taller than the cattle but now they are getting shorter and shorter, letting more light come to the grasslands of the west pastures. Therefore, some endangered plant species are now able to grow in the pasture because of valuable sunlight exposure.

One of these endangered plants is *Pulsatilla cernua*. Mount Sanbe is its major habitat. *P. cernua*, unpalatable to cattle due to its toxicity, has gradually increased its population in the grazing-revived west grassland (*Photo 14.5*) along with increases in the population of the above-mentioned *Rhododendron japonicum*.

Swertia pseudochinensis (*Photo 14.6*), another species listed in the Plant Red Data Book, has also appeared in the short-grass community close to the cattle tracks (*Figure 14.3*), while it is absent in tall *Miscanthus* grasslands outside the pasture

Photo 14.5. *Pulsatilla cernua* in the west pasture.

Photo 14.6. *Swertia pseudochinensis* in the west pasture.

area, indicating that the species cannot be conserved under the management of only annual burning. Such agricultural practices as moderately intense grazing or mowing, however, may help to conserve the species.

One of the most cost-effective ways to organize natural resource management would be to make agreements with local farmers, who, if they can spare the time, can donate their time for little cost. Therefore, instead of incurring the costs of mowing, sod cutting, and the like, the owners and managers of lands managed by cattle grazing may actually make a small profit.

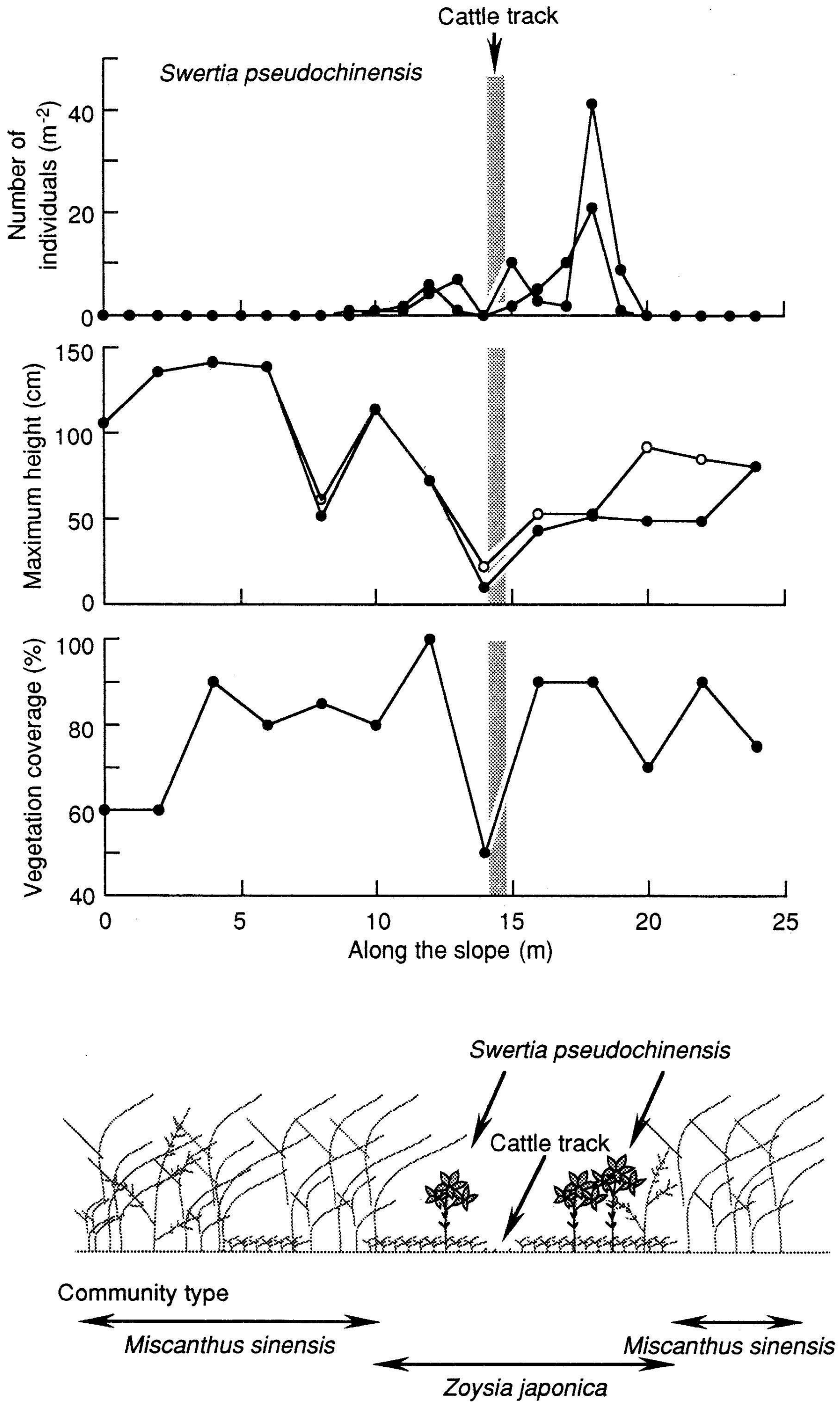


Figure 14.3. Distribution of *Swertia pseudochinensis* individuals in the west pasture.

Photo 14.7. Visitors who are fond of cattle in the west pasture.

14.4.4 Creation of amenity space with landscape animals

Being surrounded by forests, an open space occupied by grassland and turf is very valuable not only as very rare landscape, but also as an amenity space for the Japanese. In particular, the *Zoysia*-type grassland, with its short turf and grazing animals, is among those amenity spaces that are highly valued. Grazing animals are just a moving component in a grassland landscape. The scenery of the *Zoysia* grassland, with its comfortably grazing cattle and horses, instills into people feelings of a slow flow of time, leisure, and carefreeness.

In the Nishinohara pasture where cattle grazing was revived in 1996, visitors who are fond of cattle have begun to increase year by year. There is no zoo in the Shimane prefecture, so Mount Sanbe in a sense can serve as a zoo and provide the citizens an attractive space to commune with the animals (*Photo 14.7*).

Not only tourists but also community residents and neighbors appreciate and highly value the grassland landscape. According to the survey of Shoji *et al.* (1999), in the Mount Sanbe area tourists are willing to pay approximately 3,000 yen in median or 6,000 yen in mean for conservation of the grassland landscape. The total annual volume that the tourists are willing to pay amounts to 2.3 to 4.0 million yen.

14.4.5 Communication among farmers

Better and closer communication among the farmers in the Mount Sanbe area has also resulted from this change in management practices. After the grazing revival, farmers began to enlighten tourists regarding the interrelationship between cattle grazing, grassland vegetation, and the landscape. Posters regarding this topic were

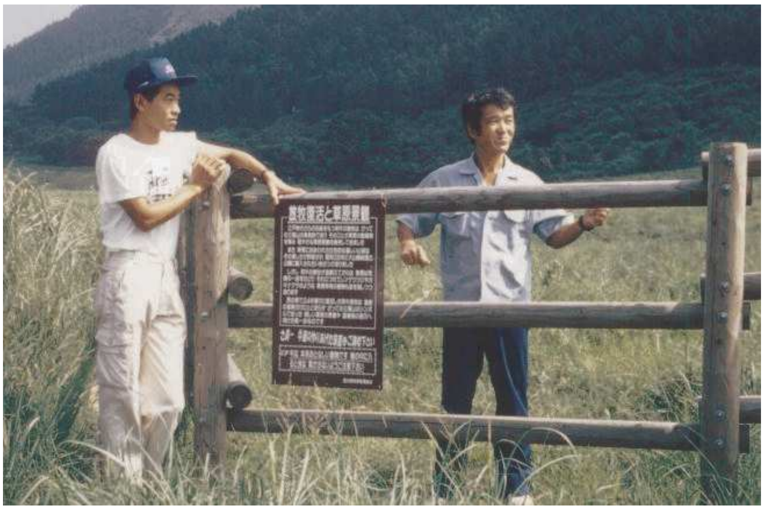


Photo 14.8. Poster to raise general awareness of grassland pasturing.

Photo 14.9. Members of the “Cattle Grazing Farmers’ Group in Oda.”

put on fences in an attempt to raise general awareness regarding grassland pasturing (*Photo 14.8*).

Stimulated by the grazing revival, the “Cattle Grazing Farmers’ Group in Oda” was founded in March 1997, with 11 member farmers, with “grazing” adopted as their motto. Since then, joint work such as bush cutting, *Zoysia* transplanting, installation of pasture fences, etc. has been carried out once a month for each farm in turn, with discussion and study meetings being held after lunch (*Photo 14.9*).

Photo 14.10. The Grassland Symposium ‘97 and Grassland Summit in Oda.

This is also an attempt to preserve and convey cattle-grazing information that has been steadily accumulated by mountain livestock farmers in the district. Farmers have started grazing their cattle in abandoned paddy fields and in forests in order to improve the environment. The introduction of cattle grazing has not only greatly reduced the labor required, but also made it possible to manage local land resources by using cattle. The number of cattle in Oda has decreased recently, but the number of members of the “Grazing Farmers’ Group” has continued to increase.

14.4.6 Closer communication with citizens

The connection to grazing has been extended to local residents; this group includes not only those who are engaged in animal production, but also those such as volunteers, citizens’ groups for nature conservation, etc. The grazing revival has made the citizens aware of the value of the grasslands and has encouraged them to take part in grassland conservation activities.

Encouraged by such a partnership between the farmers and citizens, “The Grassland Symposium ‘97 and Grassland Summit” was held in Oda city in 1996 (*Photo 14.10*). This was an event intended to increase the citizens’ consciousness. The symposium was conducted by the private sector, involving citizens, farmers and Oda city officials. Working together, people in the immediate area encouraged the growing movement to embrace a nation-wide consciousness with groups of like-minded people gathering to discuss ways to improve grassland management.

One of the annual events is the burning of the grasslands on Nishinohara. In the past, it was the city officials or workers of the forest association who did the work

Photo 14.11. Volunteers participating in the burning.

Photo 14.12. Fireguard preparation by grazing cattle.

of burning as part of their job. Now, however, those who appreciate the grassland landscape come as volunteers to help with the annual burning (*Photo 14.11*).

Since the fiscal year of 1998, preparation of the fireguard belt for annual burning has been carried out by grazing cows, in which the above-mentioned citizens' group for nature conservation has taken a leading role. It is necessary to enclose the areas (approximately 1.8 ha) with a portable electric fence to prepare the fireguard

Table 14.2. Significance of cattle grazing and problems remaining in the Mount Sanbe area.

Roles of cattle grazing	Landscape and biodiversity conservation
	Fire prevention
	Farm income support
	Employment increase
	Cultural- and rural-identity conservation
	Improvement of the rural-urban relationship
Remaining problems	Land use planning for the region
	Marketing higher-quality products and increasing tourism revenues
	Integrating conservation into viable animal production system
	Creating agro-environmental policy for sustainable development

belt (10–30 m in width) to take advantage of the cattle's "mowing ability by tongue" along the Nishinohara Pasture (*Photo 14.12*).

This endeavor proved a success when the next burning took place.

14.5 Remaining Future Problems

Table 14.2 presents the summary issues as follows. Land use planning must be carried out on a wider and more regional basis. We have to think about the tourism avenues or quality products. And we also need to integrate conservation into a variable animal production system that increases profits to the farmers, contributes to sustainable land management, favors biodiversity, and reduces the risk of fires and erosion. It is therefore important to consider the various means of policy assistance.

Furthermore, we must go further to create new demands or new markets and to facilitate closer communications with citizens. These are the secondary benefits of the improvement in the pastoral landscape in the Mount Sanbe area (*Figure 14.4*). In the case of the Mount Sanbe area, which may be spatially and conceptually differentiated from surrounding regions for its high landscape and tourism value, there are obvious market rewards for environmentally sensitive farming systems. We need to work in a holistic way to achieve these goals.

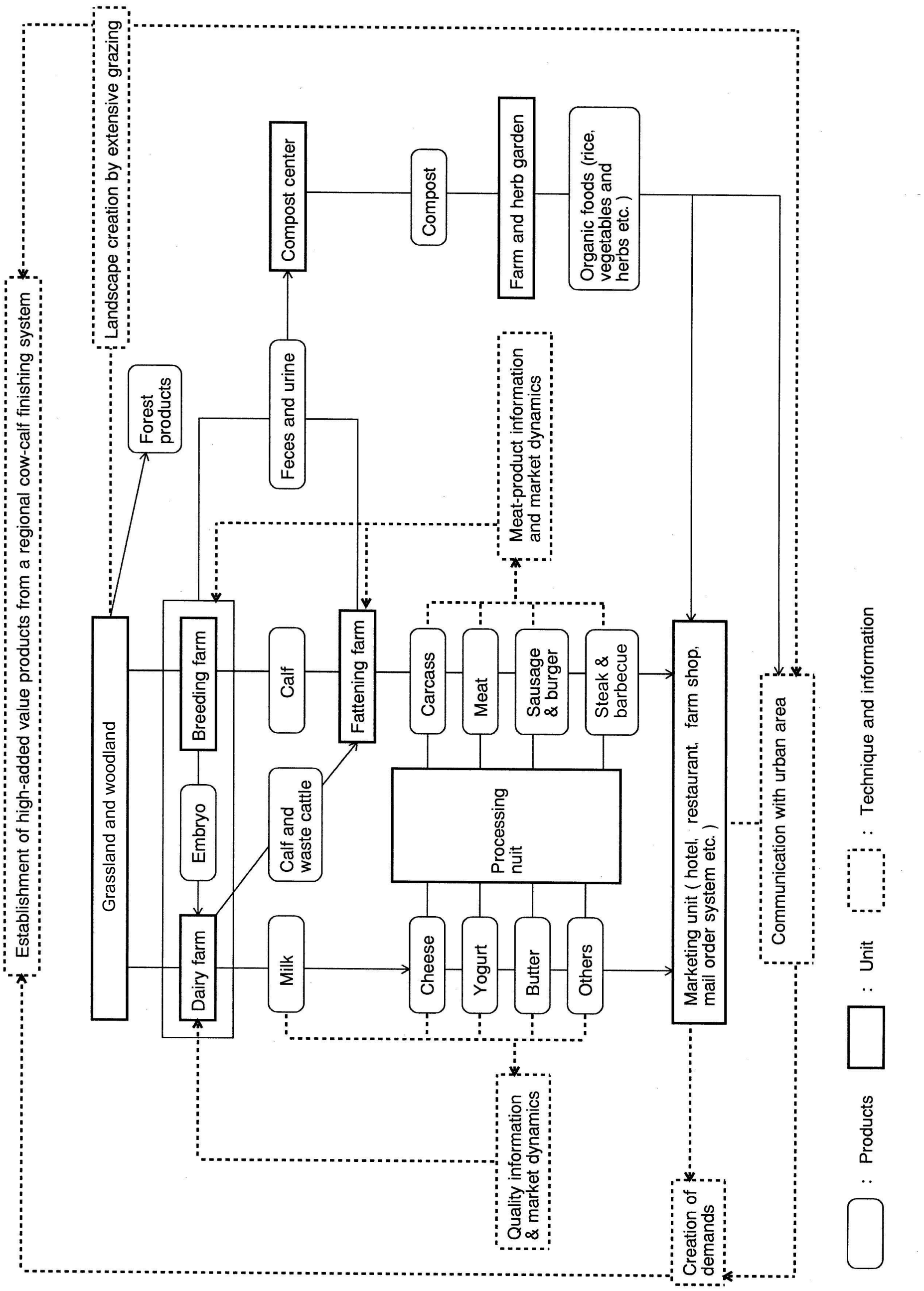


Figure 14.4. Promotion of livestock production and rural development in the Mount Sanbe area.

References

Green, B., 1996, *Countryside Conservation: Landscape Ecology, Planning and Management*, Chapman and Hall, London, UK.

Numata, M., 1994, *A concept regarding nature conservation*, Iwanamishoten, Tokyo, Japan (in Japanese).

Shoji, A., Yamamoto, Y., and Suyama, T., 1995, Analysis of landscape changes with geographic information system: A case study of Mt. Sanbe region, *Shimane Pref. J.JSIDRE*, **63**:847–853 (in Japanese).

Shoji, A., Suyama, T., and Sasaki, H., 1999, Valuing economic benefits of semi-natural grassland landscape by contingent valuation method, *Grassland Science*, **45**:88–91 (in Japanese).

Takahashi, Y., 1994, Changes in the pastures of Mt. Sanbe and remaining problems, *Bull. of Dept. of CNAES*, **H5-3**:1–39 (in Japanese).

Takahashi, Y., 1997, Grazing revival in the Mt. Sanbe area, *Bull. of Beef Cattle Science*, **63**:17–21 (in Japanese).

Takahashi, Y., and Naito, K., 1997, Conservation management for plant species diversity in a semi-natural grassland, *Shuseibutsugaku-kenkyu*, **21**:13–26 (in Japanese).

The Environment Agency, 1989, Report of vegetation survey (country issue): Third basic survey for conservation of natural environment, pp. 112–214 (in Japanese).

Chapter 15

Exploring the Relationship of Soil-Quality Index to Efficiency and Productivity Growth Measures in the Farmers' Fields

Marites M. Tiongco

Abstract

This chapter uses a data envelopment analysis (DEA) framework to estimate a soil-quality index that is consistent with the notion of technical efficiency. The estimated soil-quality index is used to identify component weights on individual soil-quality components. This chapter also demonstrates the relationship of soil quality to the inefficiency found in farmers' fields using farm level panel data from a key intensive irrigated rice area in the Philippines. The Malmquist index is also constructed using the same DEA approach. The decomposition of the Malmquist productivity index is used to reveal the soil quality components in the productivity-growth measure and thus help to separate the productivity effects of technological advances from soil quality. The empirical results are presented along with a comparative analysis of the efficiency changes with and without soil quality index.

Keywords: Data envelopment analysis, efficiency, productivity, soil-quality index, Malmquist index, intensive irrigated rice systems.

15.1 Introduction

Despite the increase in fertilizer use, the growth of rice yield in the Philippines has slowed down considerably from 3.8% in the 1970s and 1980s to –0.4% in the last decade. Studies have been undertaken on the actual and potential downward yield trends (deceleration, stagnation, and decline) on long-term experiments at the International Rice Research Institute's (IRRI's) rice stations and on some intensively irrigated rice farms. Occurrence of yield declines in farmers' fields is still in question although there are cases of farmers' stagnant or decreasing yield and productivity, which are found to be most likely due to exogenous yield shocks (Tiongco and Dawe, 1999). Recent research indicates that yield declines on long-term experiments are not widespread. In cases where yield declines occur in these intensively irrigated rice systems, these were most likely related to soil quality properties affected by soil nutrient (P and K) depletion, by prolonged submergence, and by insufficient soil drying during fallow periods (Dawe *et al*., 2000).

Soil quality has been defined as the "capability of soil to produce safe and nutritious foods and crops in a sustainable manner over the long term, and to enhance human and animal health without adversely affecting the environment." (Parr *et al*., 1992). Doran and Parkin (1994) defined soil quality as "the capacity of a soil to function within ecosystem boundaries to sustain biological productivity, maintain environmental quality, and promote plant and animal health." It is determined by complex interactions of physical, chemical, and biological soil properties, as well as human factors (Parr *et al*., 1992; Doran and Parkin, 1994). These definitions refer "soil quality" to its capacity to perform specific functions, such as the ability of the soil to function as a medium for plant growth (productivity), in the partitioning and regulation of water flow in the environment, and as an environmental buffer. In terms of agricultural production, therefore, a soil's quality refers to its ability to sustain productivity. If that is the case, then there exists a strong link between soil quality and agricultural sustainability. As soil quality increases, total factor productivity (TFP) also increases. If an agricultural system is unsustainable, it may be partly be due to the fact that soil quality is declining over time. With that hypothesis in mind, this chapter attempts to explore the relationship between soil quality and measures of efficiency and productivity.

Abundant research exists on soil quality properties in relation to farm management practices and crop yields, but there has been no generally accepted criterion for aggregating these properties into a single index. Individual soil properties may not be sufficient to quantify changing conditions so a soil quality index is desirable. A soil quality index can be an important tool in economic analysis of agricultural systems and in formulating and evaluating sustainable agricultural and environmental policies (Granatstein and Bezdicek, 1992).

This chapter, therefore, describes a method for aggregation of soil properties (biological, chemical, and physical) into a single index by employing the decomposition analysis developed by Fare *et al*. (1994). Incorporating the soil quality index in the analysis would show that soil quality is a component of the efficiency measure which is broadly defined here as a ratio of observed output to maximum potential output from a given set of inputs. It would also help separate the productivity effects of technology innovations from changes in soil quality.

Research studies of soil scientists often focus on which soil properties to include as index components but less attention has been paid to determining index weights on individual components. The presented approach also attempts to determine the index weights on individual soil properties and estimate the relative importance of each in the overall index.

Before going on with the decomposition analysis, it is important to note the two categories of soil quality indices. A soil quality index can be static or comparative-static depending on the time frame considered. A static soil quality index, S, is a function of soil quality attributes, s_i, at one point in time. The static model examines soil quality in relation to some ideal soil but does not address how soil quality may change over time. Many soil quality indicators vary significantly over the course of a growing season; the soils' physical, chemical, and biological properties change in response to use and management. A comparative-static soil quality index can be constructed by expressing soil quality, S, as a function of measurable soil attributes, si values, measuring the variation of these attributes over time (say, between time 0 and time t), and evaluating the dynamics of soil quality, dS/dt (Larson and Pierce, 1991). An aggrading soil would have a positive dS/dt and a degrading soil would have a negative dS/dt. Detecting changes in the dynamic component of soil quality is essential to evaluating the performance and sustainability of soil management systems. Hence, this chapter focuses on these two types of soil quality indices.

The next section of the chapter describes an output distance function and how it can be decomposed into components of soil quality and efficiency and productivity measures. The rest of the chapter discusses the data sources and the empirical results of the decomposition using data envelopment analysis (DEA).

15.2 Methodology

The production model considered in this chapter accounts for potentially important factors such as management inputs (like fertilizer and labor), rice produced output, and soil quality properties. At a given point in time, the production technology of the farm can be expressed through the output set, $P(\boldsymbol{x}, \boldsymbol{s})$, which represents the vector $\boldsymbol{y}$ of rice output, which can be produced using the crop input vector $\boldsymbol{x}$ and the vector of soil quality properties $\boldsymbol{s}$. That is,

$$P(\boldsymbol{x}, \boldsymbol{s}) = \{\boldsymbol{y} \varepsilon \mathbb{R} : \boldsymbol{x}, \boldsymbol{s} \text{ can produce } \boldsymbol{y}\} . \tag{15.1}$$

Soil quality here is assumed to be exogeneously predetermined and it is further assumed that the production set satisfies the axioms listed in Appendix A (Coelli *et al*., 1998). The output set $P(\boldsymbol{x}, \boldsymbol{s})$, along with the appropriate axioms, can be used to construct single-valued functions that also characterize the production technology. A generalization of the production function is the output distance function, which is defined on the output set, $P(\boldsymbol{x}, \boldsymbol{s})$, as:

$$D(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{s}) = \min\{\theta : (\boldsymbol{y}/\theta) \varepsilon P(\boldsymbol{x}, \boldsymbol{s})\} . \tag{15.2}$$

$D(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{s})$ is non-decreasing in $\boldsymbol{y}$ and increasing in $\boldsymbol{x}$, positively linearly homogeneous and convex in $\boldsymbol{y}$, and decreasing in $\boldsymbol{x}$. It takes a value of 1 if the output vector $\boldsymbol{y}$ is located on the outer boundary (or frontier) of the production possibility set. The value of $D(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{s}) \leq 1$ if $\boldsymbol{y} \varepsilon P(\boldsymbol{x}, \boldsymbol{s})$, which is also a natural measure of technical efficiency. In fact, the reciprocal of the output distance function is called the Farell output efficiency measure. As the value of the distance function increases, efficiency improves.

The concept of the output distance function in two dimensions is illustrated in *Figure 15.1*. At point A, the largest radial expansion that is on or within OPBP′ is given by the distance OB/OA. The distance AB represents technical inefficiency, and technical efficiency (TE) is OA/OB. The value of the output distance function is less than or equal to one for all points that are elements of the output set $P(\boldsymbol{x}, \boldsymbol{s})$.

Following Fare *et al*. (1994), Equation (15.2) is assumed to be multiplicatively separable in soil quality and inputs and output, and thus can be decomposed into two components:

$$D(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{s}) = S(\boldsymbol{s}) \cdot D_o(\boldsymbol{y}; \boldsymbol{x}), \tag{15.3}$$

where $S(\boldsymbol{s})$ is the natural measure of soil quality index and $D_o(\boldsymbol{y}; \boldsymbol{x})$ is an output distance function without soil quality variables. The $S(\boldsymbol{s})$ function is also nonincreasing in $\boldsymbol{s}$.

Doran and Parkin (1994) constructed soil quality indices that are nondecreasing in beneficial soil attributes. If this property holds to be desirable in a soil quality index, the reciprocal of $S(\boldsymbol{s})$ can be examined by linearly regressing it with respect to the soil attributes $\boldsymbol{s}$ to give a very restricted estimate of the individual contributions to the index. The coefficients of the parameters estimated would represent the

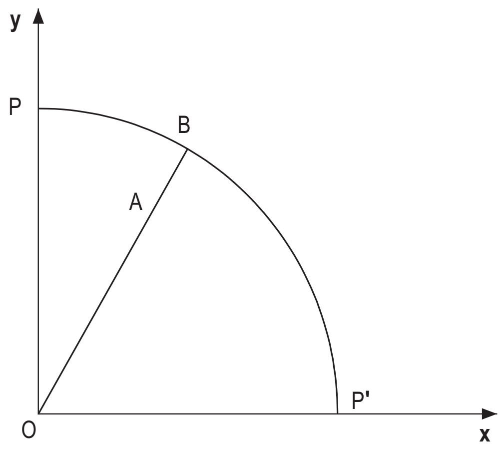


Figure 15.1. An output-oriented measure of technical efficiency.

weights on individual soil quality components. In order to isolate a theoretically preferred soil quality index, the two distance functions, $D(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{s})$ and $D_o(\boldsymbol{y}; \boldsymbol{x})$, are estimated using the data envelopment analysis framework discussed below.

15.2.1 Data envelopment analysis

DEA is a non-parametric approach which involves the use of linear programming methods to construct a piece-wise linear envelopment frontier over the data points such that all observed points lie on or below the production frontier. The model used here to calculate the output-based Farell measure of technical efficiency for each farm $k' = 1, \dots, K$, is expressed as

$$[D^*(\boldsymbol{y}^{k\prime}; \boldsymbol{x}^{k\prime}, \boldsymbol{s}^{k\prime})]^{-1} = \max \lambda^{k\prime} \,, \tag{15.4}$$

subject to

$$\lambda^{k'} y^{k'} \leq \sum_{k=1}^{K} z^k y^k \tag{15.5}$$

$$\sum_{k=1}^{K} z^k x_n^k \leq x_n^{k'} \tag{15.6}$$

$$n = 1, 2, \ldots, N$$

$$\sum_{k=1}^{K} z^k s_q^k \leq s_q^{k'} \tag{15.7}$$

$$q = 1, 2, \ldots, Q$$

$$z^k \geq 0\,, \tag{15.8}$$

where z^k refers to the weight on each specific observation.

The second distance function in Equation (15.3), $D_o^*(\boldsymbol{y}; \boldsymbol{x})$, is estimated using a similar program omitting soil quality. Once these two distance functions are estimated, for each observation, the soil quality component, $1/S(\boldsymbol{s})$ can be regressed linearly to identify the individual component weights.

The distance function also provides a natural means for aggregating all inputs and output required for measuring TFP. Caves *et al*. (1982) showed an output-oriented Malmquist index which is the ratio of distance functions from two time periods. Fare *et al*. (1994) developed a Malmquist-type index that can be represented as the geometric mean of two distance function ratios, where both the technology and the inputs and output are time-dated.

15.2.2 The Malmquist productivity index

Fare *et al*. (1994) decomposed productivity growth into two mutually exclusive components: technical change and efficiency change over time. They calculated productivity change as the geometric means of two Malmquist productivity indices for year t and $t+1$ reference technologies:

$$M_o\left(y^{t+1}, x^{t+1}, y^t, x^t\right) = \left(\frac{D^t\left(y^{t+1}, x^{t+1}\right) D^{t+1}(y^{t+1}, x^{t+1})}{D^t\left(y^t, x^t\right) D^{t+1}\left(y^t, x^t\right)}\right)^{1/2}. \tag{15.9}$$

This expression can be factored into the product of technical change and efficiency change:

$$M_o\left(y^{t+1}, x^{t+1}, y^t, x^t\right) = \frac{D^{t+1}\left(y^{t+1}, x^{t+1}\right)}{D^t\left(y^t, x^t\right)} \left(\frac{D^t\left(y^{t+1}, x^{t+1}\right) D^t\left(y^t, x^t\right)}{D^{t+1}\left(y^{t+1}, x^{t+1}\right) D^{t+1}\left(y^t, x^t\right)}\right)^{1/2}. \tag{15.10}$$

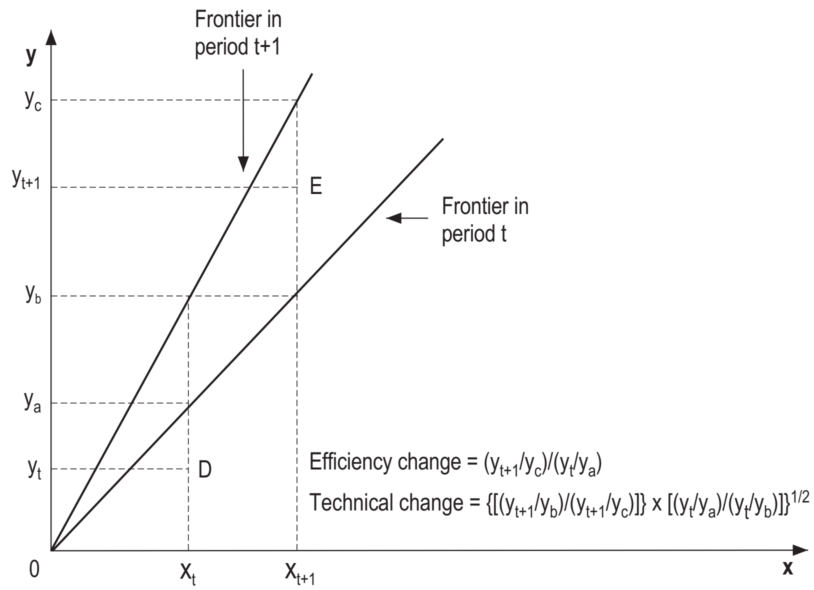


Figure 15.2. Two components of the Malmquist productivity index.

The ratio outside the brackets measures the change in relative efficiency (i.e., the change in the distance of observed production from maximum feasible production) between years t and $t + 1$, while the bracketed term measures the shift in technology (or technical change) between the two periods evaluated at x^t and x^{t+1}. The decomposition of the Malmquist productivity index allows us to identify the contributions of catching up in efficiency and innovation in technology to the TFP growth. According to Fare *et al*. (1994), a Malmquist index greater than one reveals improved productivity. Malmquist indices less than one indicate a decline in productivity. In addition, improvements in any of the two components of the Malmquist productivity index are also associated with values greater than one, and declines are associated with values less than one.

To illustrate the decomposition graphically, consider a single input x producing a single output y operating under constant returns to scale (see *Figure 15.2*). The farm produces at the points D and E in periods t and $t + 1$, respectively. In each period, the farm is operating below the technology frontier; hence there is technical inefficiency in both periods. Using Equation (15.10), we have:

Efficiency change = $(y_{t+1}/y_c)/(y_t/y_a)$
Technical change = $\{[(y_{t+1}/y_b)/(y_{t+1}/y_c)] \times [(y_t/y_a)/(y_t/y_b)]\}^{1/2}$.

Following the above methodology, an enhanced decomposition of the Malmquist index is used in this chapter to analyze the productivity growth and soil quality changes in agricultural production. To construct the Malmquist productivity index of farm k' between t and $t+1$, the DEA approach is used to calculate the following distance functions: $D^{*t}(y^t, x^t, s^t)$, $D^{*t+1}(y^t, x^t, s^t)$, $D^{*t}(y^{t+1}, x^{t+1}, s^{t+1})$, $D^{*t+1}(y^{t+1}, x^{t+1}, s^{t+1})$, $D_o^t(y^t, x^t)$, $D_o^{t+1}(y^t, x^t)$, $D_o^t(y^{t+1}, x^{t+1})$, and $D_o^{t+1}(y^{t+1}, x^{t+1})$. These distance functions are computed by comparing observations in one time period with the best-practice frontier of another time period, assuming constant returns to scale technology. The Malmquist-type index is defined as

$$M^*\left(y^{t+1}, x^{t+1}, s^{t+1}, y^t, x^t, s^t\right)$$

$$= \left(\frac{D^{*t}\left(y^{t+1}, x^{t+1}, s^{t+1}\right) D^{*t+1}\left(y^{t+1}, x^{t+1}, s^{t+1}\right)}{D^{*t}\left(y^t, x^t, s^t\right) D^{*t+1}\left(y^t, x^t, s^t\right)}\right)^{1/2} \tag{15.11a}$$

$$= \left(\frac{S^t(s^{t+1}) D_o^{*t}\left(y^{t+1}, x^{t+1}\right) S^{t+1}(s^{t+1}) D_o^{*t+1}\left(y^{t+1}, x^{t+1}\right)}{S^t(s^t) D_o^{*t}\left(y^t, x^t\right) S^{t+1}(s^t) D_o^{*t+1}\left(y^t, x^t\right)}\right)^{1/2} \tag{15.11b}$$

$$= \left(\frac{S^t(s^{t+1}) S^{t+1}(s^{t+1})}{S^t(s^t) S^{t+1}(s^t)}\right)^{1/2} \times M_o\left(y^{t+1}, x^{t+1}, y^t, x^t\right) . \tag{15.11c}$$

The first term in (15.11c) denoted by $\Delta S(s_t, s_{t+1})$ is the comparative-static soil-quality index, which can be used as dependent variable in a linear regression to identify the proportional changes in soil quality component weights. The index M_o is a modified Malmquist productivity index without soil-quality variables.

15.3 Data Sources and Descriptions

The data used in this chapter are on-farm monitoring data on intensively irrigated rice farms collected by the Reversing Trends of Declining Productivity (RTDP) project of the IRRI. The RTDP project or the mega project was initiated in 1994 (and is still on-going at present) in key irrigated rice domains in Asia. Its intent is to identify the major constraints to productivity of intensive rice-rice systems and to obtain the baseline of long-term trends in productivity and soil quality.

Phase I (1994–1996) of the mega project monitored detailed biophysical and socioeconomic indicators at five sites with a tropical climate (Central Luzon in the Philippines, Tamil Nadu in South India, Mekong Delta in South Vietnam, Central Thailand, and West Java). Phase II (1997–2000) expanded the work to three new sites with a subtropical climate (Southeast China, North Vietnam, and North India). These are the first long-term micro-level data that monitor both socioeconomic and biophysical variables in key irrigated areas in South and East Asia.

For the purpose of this chapter, the Central Luzon data on socioeconomic and biophysical variables are used to evaluate the relationship between TFP measures and soil quality index. The data collection started in the 1994 wet season (WS) with 47 farmers and was reduced to 27 farmers in the 1997 and 1998 WS. A total of 135 farmers belonged to the balanced panel data for the WS, and 72 farmers for the dry season (DS). The number of farmers in the dry season was smaller because only 19 farmers planted rice during the 1997 DS.

Data collected on soil variables include carbon (C); nitrogen (N); cation exchange capacity (cec); negative log of hydronium ion (pH); exchangeable potassium (K); phosphorous (P); particle-size distribution, e.g., silt, clay, and sand; exchangeable calcium (CA); exchangeable magnesium (Mg); exchangeable sodium (Na); electrical conductivity; and total cations. Until recently, there has been no definite soil quality index that has been accepted by soil scientists, however, the soil data set of the mega project represents a potential subset of the soil quality indicators proposed by several soil-quality researchers. The soil variables were analyzed for correlation before including them in the decomposition analysis. Only variables that are not highly correlated with each other were included. The reduced data set that comprises $\boldsymbol{s}$, the vector of soil-quality, are carbon-nitrogen ratio (C/N) as biological indicator; P, K, and pH as chemical indicators; and percent of silt particles in the soil (% silt) as physical indicator. The greatest degree of correlation among the elements of $\boldsymbol{s}$ is the interaction between silt and pH, which has a correlation coefficient of 0.39.

Socioeconomic variables include specific whole farm data on rice area planted, grain yield, labor, and material inputs for all crop management operations such as land preparation, crop establishment, fertilizer application, pest control, irrigation, harvest, and post-harvest activities.

The input vector $\boldsymbol{x}$ includes the nitrogen component of fertilizer in kilograms per hectare (kg ha^{-1}); insecticide use in grams per active ingredient (gm ai^{-1}); seeds (kg ha^{-1}); total labor, which includes family and hired labor in days per hectare (d ha^{-1}); and machine rental in pesos per hectare (pesos ha^{-1}), deflated with consumer price index. The output vector is rice production in kg ha^{-1} (yield kg ha^{-1}).

Table 15.1. Average values of soil properties and socioeconomic variables, DS 1995–1998 and WS 1994–1998.

	DS 1995–1998		WS 1994–1998	
Variables	Mean	Std Dev	Mean	Std Dev
Soil quality variables				
Carbon/nitrogen ratio	9.64	1.41	9.41	1.35
Exchangeable K (cmol kg^{-1})	0.07	0.06	0.08	0.07
pH	5.99	0.24	6.02	0.22
Phosphorus (mg kg^{-1})	4.26	3.80	11.33	13.21
% silt	36.44	23.72	42.52	21.17
Socioeconomic variables				
Nitrogen (kg ha^{-1})	187.68	101.58	113.35	67.89
Insecticide (gm ai^{-1})	69.03	183.08	186.45	484.51
Machine rental (pesos ha^{-1})	542.41	862.89	542.80	745.31
Seed (kg ha^{-1})	202.54	82.91	224.24	154.20
Total labor (d ha^{-1})	27.52	25.00	35.76	28.89
Yield (kg ha^{-1})	7,037.67	1,139.33	4,182.58	1,424.65

Table 15.2. Static efficiency decomposition by year and season.

Year/Season	$D(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{s})$	$D_o(\boldsymbol{y}; \boldsymbol{x})$	$S(\boldsymbol{s})$	$1/S(\boldsymbol{s})$
WS (N=135)				
1994	0.687	0.540	1.425	0.816
1995	0.893	0.778	1.223	0.869
1996	0.765	0.617	1.351	0.823
1997	0.888	0.751	1.366	0.844
1998	0.863	0.797	1.114	0.919
DS (N=72)				
1995	0.941	0.833	1.185	0.870
1996	0.957	0.807	1.337	0.850
1997	0.923	0.813	1.273	0.877
1998	0.961	0.860	1.159	0.899

The data set provided above is used to estimate the DEA-based distance functions in Equations (15.3) and (15.11), and to approximate the weights in a static and comparative-static soil quality index. *Table 15.1* summarizes the average values of the socioeconomic and soil quality variables used in the analysis.

15.4 Results and Discussion

The output-based distance functions in Equation (15.3) were estimated using the DEA program developed by Coelli (1996). *Table 15.2* summarizes the values of

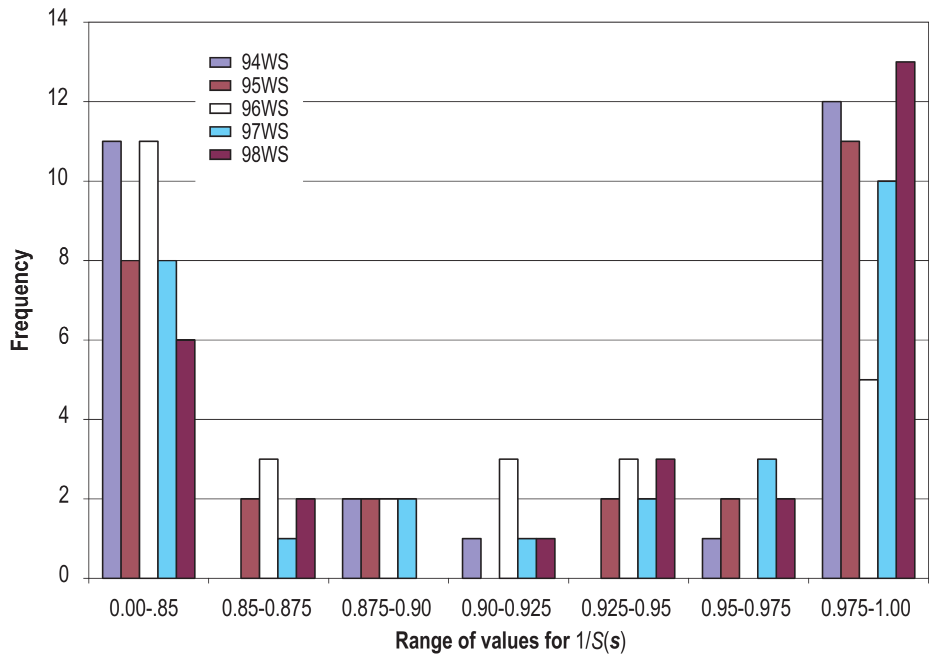


Figure 15.3. Histogram of $1/S(\boldsymbol{s})$, 1994–1998 WS.

$D(\boldsymbol{y};\boldsymbol{x},\boldsymbol{s})$, $D_o(\boldsymbol{y};\boldsymbol{x})$, $S(\boldsymbol{s})$, and $1/S(\boldsymbol{s})$. On average, technical efficiency for the 27 farmers during the WS is understated by 21.4% in 1994, 12.9% in 1995, 19.3% in 1996, 15.4% in 1997, and 7.6% in 1998 when measured by $D_o(\boldsymbol{y};\boldsymbol{x})$ instead of $D(\boldsymbol{y};\boldsymbol{x},\boldsymbol{s})$. Similarly, for the DS, the technical efficiency for the 18 farmers is understated by 11.5% in 1995, 15.7% in 1996, 11.9% in 1997, and 10.5% in 1998. Thus, accounting for soil quality helps explain, on the average, about 30–55% (for the WS) and about 59–78% (for the DS) of the inefficiency found in the individual farms. The means were also calculated and tested whether the difference between the two groups (with and without soil quality variables) was statistically significant. The results show that the difference between the technical efficiencies measured by $D_o(\boldsymbol{y};\boldsymbol{x})$ and $D(\boldsymbol{y};\boldsymbol{x},\boldsymbol{s})$ are statistically significant at 1% level of significance. *Figure 15.3* and *Figure 15.4* show the distribution of $1/S(\boldsymbol{s})$ over the range 0 to 1.

The correlation coefficients of the yield and the soil quality index are estimated. Relatively, there is some positive correlation between yield and soil quality index, which implies that those farms with high soil quality indices correspond to high yields. Within the data set, the only positive correlation that was statistically significant (5% level of significance) occurred during the 1997 WS (0.36). The other seasons have positive correlation coefficients but none that are statistically significant.

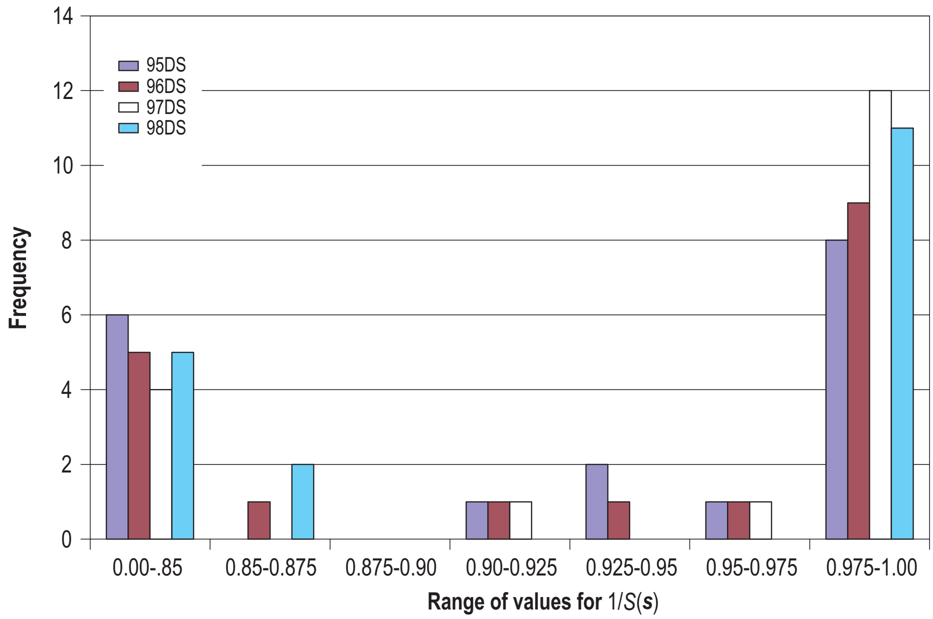


Figure 15.4. Histogram of $1/S(\boldsymbol{s})$, 1995–1998 DS.

The Malmquist productivity change index and its components were also computed using DEAP 2.0 formulated by Coelli (1996). The various Malmquist index measures were computed from optimal solutions for each of 27 farms and for every consecutive pair of years (1994/95, 1995/96, 1996/97, 1997/98). As shown in *Table 15.3*, a value of the comparative-static soil quality indices $(1/\Delta S(s_t, s_{t+1}))$ of less than one corresponds to a deterioration in soil quality for the period under study. Since the data used generated only 4 points of change over time, conclusions cannot easily be drawn as to whether there exists a declining trend in the soil quality index. These changes in soil attributes can form the basis for measures of sustainable development if there is at least 10 years of data.

TFP growth for the WS (using the data set with soil quality variables) as a whole increased from 1994–1998, which reveals growth during these periods (*Table 15.4*). Average change in productivity over the period was 1.1% annually. On average, technical change increased productivity by 2.4% per year which offset the decline of efficiency by an average of 1.2%. Similarly for the DS (using the data set with soil quality variables), an improvement in TFP is revealed during the time period. This is mainly due to the average increase of technical change by 11.8%, which offset the decline of efficiency change by an average of 1.6%. To augment technical progress, given the limited opportunities to expand cultivable land, the greater potential lies in increasing or attracting investment in agricultural research

Table 15.3. Annual change in soil-quality index by season.

Year/Season	$\Delta S(s_t, s_{t+1})$	$1/\Delta S(s_t, s_{t+1})$
WS (N=135)		
1994/1995	1.032	0.969
1995/1996	1.134	0.882
1996/1997	1.069	0.935
1997/1998	1.200	0.833
Mean	1.107	0.903
DS (N=72)		
1995/1996	1.340	0.746
1996/1997	1.339	0.747
1997/1998	1.229	0.814
Mean	1.302	0.768

Table 15.4. Average annual changes of Malmquist indices (with and without soil properties) by season.

	With soil properties			Without soil properties		
Year/Season	Malmquist index	Technical change	Efficiency change	Malmquist index	Technical change	Efficiency change
WS (N=135)						
1994/1995	0.873	1.126	0.775	0.846	0.794	1.065
1995/1996	0.915	0.656	1.396	0.807	0.626	1.288
1996/1997	1.092	1.266	0.863	1.021	1.486	0.687
1997/1998	1.199	1.174	1.022	0.999	1.016	0.983
Mean	1.011	1.024	0.988	0.913	0.931	0.981
DS (N=72)						
1995/1996	1.040	1.002	1.038	0.776	0.813	0.954
1996/1997	1.165	1.405	0.829	0.870	1.201	0.725
1997/1998	1.154	1.042	1.108	0.939	0.665	1.413
Mean	1.118	1.136	0.984	0.859	0.866	0.992

and technical development (such as better irrigation facilities or new seeds) in rice production. No improvement was seen when the data set without soil properties was analyzed.

Since the Malmquist productivity index and its components are multiplicative, we can calculate the cumulated Malmquist productivity index and its components such as the cumulated technical change index, the cumulated efficiency change index, and the cumulated soil quality change index. The cumulated indices measure the total changes in TFP, technical efficiency, soil quality, and technology over the 1994–1998 WS only. To see the patterns of the changes in these indices, we plot the cumulated productivity indices against time (with soil quality indicators), as

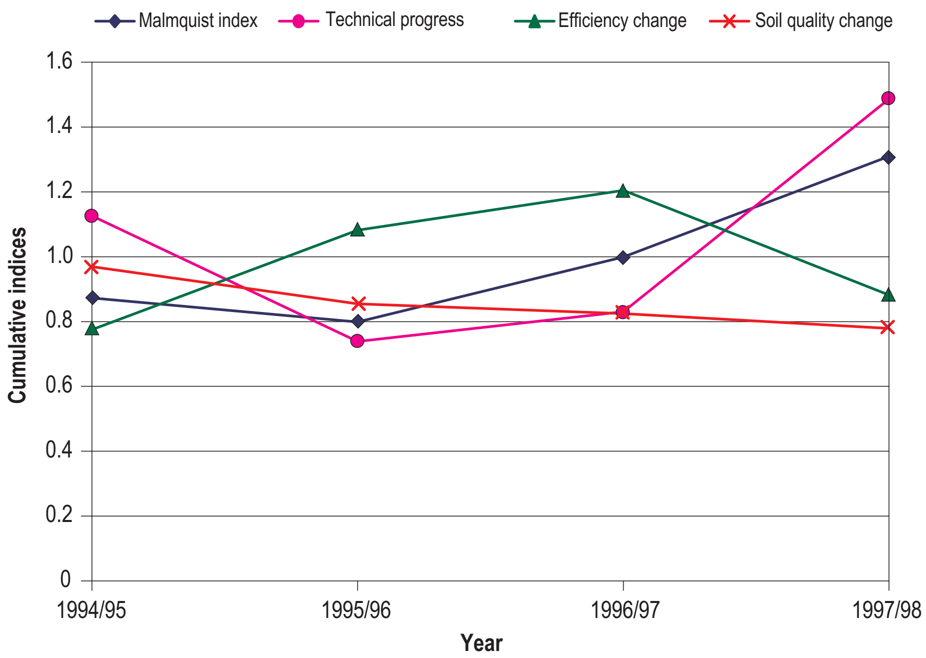


Figure 15.5. Cumulative indices (with soil quality variables), 1994–1998 WS.

shown in *Figure 15.5*. TFP exhibits an increasing trend. The soil quality trend is deteriorating. This decline could be due to inadequate application of fertilizers, decline in soil nutrient capacity, or continuous soil puddling which destroys soil structure. This finding, however, has to be reviewed later since the time period of study is quite short.

Table 15.5 presents the results of the ordinary least squares (OLS) regressions when the static soil-quality index is estimated with respect to each individual soil attribute for a specific time period (1995–1996). The choice of a data sample is for illustrative purposes only. The relative importance of the soil attributes in the soil-quality index is given $\beta_q(\sigma_q/\sigma_s)$, where β_q is the estimated coefficient, σ_q is the standard deviation of the attribute, and σ_s is the standard deviation of $1/S(\boldsymbol{s})$. The results suggest that for the wet season, only the coefficient of C/N is significant at 1%. It plays the strongest role in the index relative to the other attributes and should be given a linear weight of 0.06. Plant growth is highly affected by C/N ratio due to the release of N for uptake. For the dry season, coefficients of C/N and pH are statistically significant at 1% and 10%, respectively. The carbon/nitrogen ratio must be given a linear weight of 0.74, which is more than twice the next largest weight 0.28 for pH. The mean value of pH for the data set used in this analysis is near neutral (6–6.02), where the essential nutrients N, P, and K are abundant

Table 15.5. OLS estimates of $1/S(\boldsymbol{s})$, 1995–1996.

Soil Quality (s_q)	β_q	$\beta_q(\sigma_q/\sigma_s)$	t-stat
Wet season 1995			
Carbon/nitrogen	0.104	0.060	2.383[a]
P	0.010	0.001	0.459
Exchangeable K	0.292	0.474	0.734
Silt	–0.004	7.6E–05	–1.118
pH	0.119	0.079	0.787
Constant	–0.936		–0.818
R^2	0.12		
Dry season 1996			
Carbon/nitrogen	0.142	0.743	4.149[b]
P	0.013	0.245	1.174
Exchangeable K	0.075	0.003	0.028
pH	0.291	0.277	1.857[c]
Silt	0.0002	0.018	0.094
Constant	–2.270		–2.354[a]
R^2	0.67		

[a] 5% level of significance.
[b] 1% level of significance.
[c] 10% level of significance.

and easily absorbed by the crop. The linear weights for the comparative-static soil quality index are not shown here because one of the attributes, % silt, contributed to the singularity of the matrix when regressed with other soil attributes.

15.5 Summary and Concluding Comments

This chapter reconciles the notion of soil quality index with efficiency and productivity using the DEA framework. The measures of technical efficiency when soil quality variables are included, have demonstrated higher estimates as compared to when only socioeconomic variables are included. The addition of an extra input or output in a DEA model cannot result in a decrease in the efficiency scores (Coelli *et al*., 1998). Hence, when soil variables were added, the efficiency estimates were higher when compared to the estimates without soil variables. However, it is the *choice* of the soil variables that matters because it affects the magnitude of the efficiency scores. Thus, the soil quality variables included in the analysis help to explain the inefficiency found in individual farms. To further explain the role of individual soil quality properties, the OLS technique finds that C/N and pH should be given the greatest weights in a very restricted linear estimate of the soil quality index.

The decomposition analysis of soil quality, efficiency, and productivity measures showed a deterioration in the soil quality trend. This trend implies that under an intensively irrigated rice-rice cropping system, soil quality is decreasing. Hence, improved nutrient management practices in farmers' fields should be implemented and monitoring of the changes of indigenous nutrient supply should continue for a longer period of time. The data set shows that the deteriorating soil quality is offset by technical improvement that resulted to an increasing trend in TFP growth. These results indicate that technical change was the dominant force augmenting productivity growth during this period. Therefore, policies must be designed to encourage technical progress and must be accompanied by successful technology diffusions (e.g., new seeds).

The comparative static soil quality index can be a basis for measuring sustainable management, just as the TFP indices are. The robustness of the results in this chapter could be further investigated if the analytical time period covered at least 10 years of data. It should also be noted that the results presented here are site-specific to the farm fields and practices of Central Luzon in the Philippines.

Acknowledgment

The author appreciates helpful comments from Dr. David Dawe on an earlier draft of this chapter.

References

Caves, C. W., Christensen, L. R., and Diewert, W.E., 1982, Multilateral comparisons of output, input, and productivity using superlative index numbers, *Economic Journal*, **92**:73–86.

Coelli, T., 1996, A Guide to DEAP Version 2.1: A DEA (Computer) Program, Centre for Efficiency and Productivity Analysis, Working Paper 96/08, Department of Econometrics, University of New England, Armidale, Australia.

Coelli, T., Rao, P.D.S., and Battese, G., 1998, *An Introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, Boston, MA, USA, Dordrecht, Netherlands, and London, UK.

Dawe, D. *et al*., 2000, How widespread are yield declines in long-term rice experiments in Asia?, *Field Crops Research*, **66**:175–193.

Doran, J.W., and Parkin, T.B., 1994, Defining and assessing soil quality, in J.W. Doran, D.C. Coleman, D.F. Bezdicek, and B.A. Stewart, eds, *Defining Soil Quality for a Sustainable Environment*, Spec. Publ. 35, ASA and Soil Science Society of America, Madison, WI, USA, pp. 3–21.

Fare, R., Grosskopf, S., Norris, M., and Zhang, Z., 1994, Productivity growth, technical progress, and efficiency change in industrialized countries, *American Economic Review*, **84**:66–83.

Farrell, M.J., 1957, The measurement of productive efficiency, *J. of the Royal Statistical Society*, **120**:253–282.

Granatstein, D., and Bezdicek, D.F., 1992, The need for a soil quality index: Local and regional perspectives, *Am. J. Altern. Agric.*, **17**:12–16.

Larson, W.E., and Pierce, F.J., 1991, Conservation and enhancement of soil quality, in *Proceedings of the International Workshop on Evaluation for Sustainable Land Management in the Developing World*, Vol. 2, IBSRAM Proc. 12(2), International Board for Soil Research and Management, Bangkok, Thailand.

Parr J.F., Papendick, R.I., Hornick, S.B., and Meyer, R.E., 1992, Soil quality: Attributes and relationship to alternative and sustainable agriculture, *Amer. J. of Alter. Agri.*, **7**:5–11.

Tiongco, M., and Dawe, D., 1999, Is Green Revolution Sustainable? Long-term Productivity Trends in a Sample of Philippine Rice Farms, Poster paper presented at the 24th International Conference of Agricultural Economists, 13–18 August, Berlin, Germany.

Appendix

The properties of the output set $P(\boldsymbol{x}, \boldsymbol{s})$ are as follows. For each $\boldsymbol{x}$, the output set $P(\boldsymbol{x}, \boldsymbol{s})$ is assumed to satisfy:

1. $0 \varepsilon P(\boldsymbol{x}, \boldsymbol{s})$: nothing can be produced out of a given set of inputs (that is, inaction is possible);
2. Non-zero output levels can't be produced from zero level of inputs;
3. $P(\boldsymbol{x}, \boldsymbol{s})$ satisfies strong disposability of outputs: if $y \varepsilon\ P(\boldsymbol{x}, \boldsymbol{s})$ and $\boldsymbol{y}^* \leq \boldsymbol{y}$ then $\boldsymbol{y}^* \varepsilon\ P(\boldsymbol{x}, \boldsymbol{s})$. Alternatively, weak disposability states that if a vector of outputs, $\boldsymbol{y}$, can be produced from a given of input vector $\boldsymbol{x}$ then any contraction of $\boldsymbol{x}$, λy, of $0 < \lambda < 1$, can also be produced with x.
4. $P(\boldsymbol{x}, \boldsymbol{s})$ satisfies strong disposability of inputs: if y can be produced from x, then y can be produced from any $x^* \geq x$.
5. $P(\boldsymbol{x}, \boldsymbol{s})$ is closed;
6. $P(\boldsymbol{x}, \boldsymbol{s})$ is bounded;
7. $P(\boldsymbol{x}, \boldsymbol{s})$ is convex.

Part IV: Water Management

Chapter 16

Climate Change Impacts on Water-Stressed Agriculture in Northeast China

Günther Fischer and David Wiberg

Abstract

Both economic and environmental factors influence land planning decisions. The suitability of land for agriculture, for instance, depends upon climate and soil conditions as well as farm management. In China, where arable land is scarce and declining due to urban growth and the population is close to 1.3 billion and increasing, land planning is of increasing importance. Water resource availability is also a key concern throughout China, but particularly in the Northeast where much of the country's agriculture resides. Water could well become a limiting factor in expanding agriculture and industry in this region.

In this chapter, two complementary models are used to analyze the impacts of scenarios from three general circulation models on agricultural potential in the North China Plain. LUC's implementation of FAO's agro-ecological zones (AEZ) methodology is employed to assess crop production potential in the region, and a runoff model developed within LUC called CHARM is then used to calculate changes in runoff. The runoff generated by the CHARM model can be compared with AEZ results to indicate whether enough water is available to satisfy the irrigation demand generated by the cropping patterns evaluated by the AEZ methodology for the different climate scenarios.

Analysis of the impacts of the climate scenarios using AEZ and CHARM indicate a generally favorable response of these regions to climate change. Under all three scenarios for 2050, cereal production potentials increase by between 10 and 15%. In some scenarios, however, the increase is only achieved by substantial increases in irrigation water. Water resources in the region are already too scarce to meet today's demand. Although several scenarios produce increased runoff, demand would still not be met without importing water from other regions of China. Furthermore, climate variability is expected to increase under all scenarios resulting in increasing frequency of extreme events.

Keywords: Climate change, hydrology, agro-ecological zones (AEZ), North-China Plain, water scarcity, irrigation, climate variability.

16.1 Introduction

Various approaches for analyzing and projecting regional impacts of global environmental change have been developed at the International Institute for Applied Systems Analysis (IIASA). Initially these studies were geared toward assessing the potential impacts of global climate change (Rosenzweig and Parry, 1994; Fischer *et al*., 1996; Fischer, forthcoming).

More recently, changes in land use and land cover were recognized as being central to the study of global environmental change (Turner *et al*., 1995). In addition to their cumulative long-term global dimensions and their potential responsiveness to global environmental change, such alterations can have profound regional environmental implications during the life span of current generations, including various themes central to the debate of sustainability such as reduced biodiversity, changed land productivity, climate feedback, robustness of the land use systems with respect to economic or environmental shocks, or increasing water scarcity problems, e.g., the lowering of groundwater tables.

The importance of these topics, and the need for innovative and interdisciplinary approaches to studying the nature of land-use and land-cover changes, have prompted IIASA to establish its project on Modeling Land-Use and Land-Cover Changes in Europe and Northern Asia (LUC). The project has two main goals: first and foremost, to develop new concepts that address the methodological challenges of projecting complex human-environment systems; and second, to apply these concepts within regional assessments of economically viable options for land use and food policy. The current application focus has been on China.

16.1.1 Model-based analysis of land and water resources

Land use and food systems represent a critical intersection of the economy and the environment. While studies of the Earth system are concerned with land-cover changes and alterations of biochemical cycles of carbon, nitrogen, etc., social science disciplines and political attention relate to food security, rural development, and sustainability of land-use systems. Land-use changes are most often directly linked with economic decisions. This recognition has led LUC to choose an economic framework as its organizing principle, resulting in a broad set of project activities geared towards providing a biophysical and geographical underpinning to the representation of actors and land-based economic sectors in modeling land and water use decisions.

LUC has been aiming to contribute to the complex and broad research theme with theoretically sound yet practical new approaches, including integration of diverse statistical and geographical data sets within a Geographical Information System (GIS), agro-ecological zones (AEZ) assessments of environmental constraints and land productivity potentials, analytical tools for water resources assessments, and development of decision tools for evaluating policy options concerning land use and agriculture.

Three complementary approaches in land-use analysis were applied within LUC's study of China's food and land-use prospects (*Figure 16.1*). First, an enhanced AEZ assessment model is used to provide a spatially explicit measure of crop suitability and land productivity potentials for rain-fed and irrigated conditions. This provides inputs to scenario analysis, based on an extended input-output model, to quantify future land requirements associated with the projected demographic trends and economic activities (see Chapter 12 by Hubacek and Sun, in this volume). Finally, changes in land and water use are viewed as dependent on how these resources are transformed and managed by human activity. The underlying decision problem is cast in the form of a welfare optimum model to elaborate socially desirable and economically efficient trajectories of resource uses and transformations. For this purpose, special emphasis was placed on specifying a set of production relations of the land-based economic sectors such that these would permit the integration of spatially varied biophysical factors into the economic model.

Global environmental change issues are long-term and are related to questions of resource development and investment planning. Current demographic and socioeconomic trends suggest that the next 30 to 50 years will be decisive for managing viable transitions towards sustainable land-use systems. The LUC project is therefore concentrating its analysis on the period up to 2050. In this chapter we concentrate on a specific subset of questions of importance in the discussion of China's food system prospects. Geographically, we look at food production systems in three river basins of North China encompassing China's breadbasket,

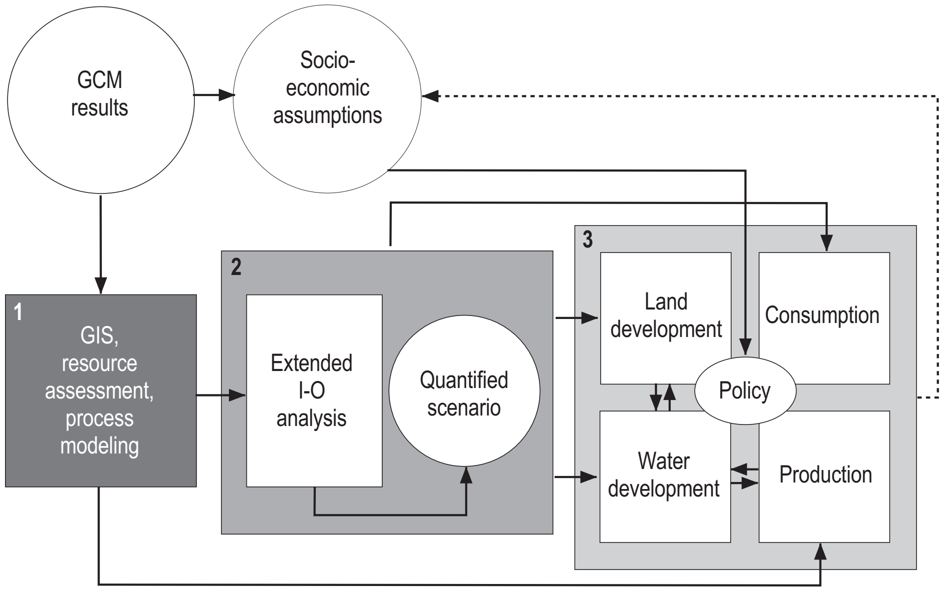


Figure 16.1. Scheme of complementary methodologies in LUC China study. Note: 1 = biophysical assessment; 2 = extended input-output analysis; 3 = intertemporal welfare optimization model.

the North-China Plain (NCP). Thematically, we limit the analysis to the question of if and when alterations in production conditions due to global environmental change may exert additional stress on these systems, which already suffer from water stress under current conditions. Hence, we present some background and results from two biophysical assessment models, the AEZ and CHARM (**C**limate and **H**uman-**A**ctivities sensitive **R**unoff **M**odel) models.

16.1.2 Crop suitability and land productivity assessment based on AEZ modeling

LUC has recently completed a new implementation of a series of land evaluation steps, originating from the Food and Agriculture Organization of the United Nations (FAO) and widely known as the agro-ecological zones (AEZ) methodology (for details of the AEZ approach see http://www.iiasa.ac.at/Research/LUC/GAEZ). In applying this methodology to the territory of the former Soviet Union, Mongolia, and China, LUC has extended the AEZ from its original focus on tropical and subtropical conditions to the seasonal temperate and boreal climate zones. The system works with recent digital databases to quantify agricultural production risks as expressed by historical variability and considers possible impacts of climate change on the prevalence of constraints to crop production and of agricultural potentials. Another objective was to generate geographically explicit information that could be embedded in LUC's economic analysis and that would allow consistent linkage to water availability assessments.

The choice of applying the principles of the AEZ methodology (FAO, 1984, 1985; UNDP/SSTC/FAO/SLA, 1994) within the land productivity component of LUC is based on the fact that AEZ follows an environmental approach. It provides a geographic framework for establishing a spatial inventory and database on land resources and crop/grassland production potential. The data requirements are sufficiently limited to enable full coverage of a country or larger region, and it makes maximum use of readily available data. Moreover, it is comprehensive in terms of coverage of factors affecting production. In its simplest form, the AEZ framework contains three elements:

1. Selected agricultural production systems with defined input/output relationships, and crop-specific environmental requirements and adaptability characteristics [Land Utilization Types (LUT)].
2. Geo-referenced land resources data (climate, soil and terrain data).
3. Models for the calculation of potential yields and procedures for matching crop/LUT environmental requirements with the respective environmental characteristics contained in the land resources database, by land unit and grid-cell.

Agro-ecological zoning involves the inventory, characterization, and classification of the land resources, to enable assessments of the potential of agricultural production systems. This characterization includes all components of climate, soils, and landform, which are basic for the supply of water, energy, nutrients, and physical support to plants.

A water-balance model is used to quantify the beginning and duration of the period when sufficient water is available to sustain crop growth. Soil moisture conditions together with other climate characteristics (e.g., radiation and temperature) are used in a simple crop growth model to calculate potential biomass production and yield. This *potential* yield is then combined in a semi-quantitative manner with a number of reduction factors directly or indirectly related to climate (e.g., pests and diseases) and soil and terrain conditions. The reduction factors, which are successively applied to the potential yields, vary with crop type, the environment (in terms of climate, soil, and terrain conditions), and assumptions on level of input/management. The final results consist of *attainable* crop yields under sets of pre-defined standardized production circumstances, referred to as LUTs.

Results have been classified in five basic suitability classes according to attainable yield: VS - very suitable (80–100% of maximum attainable yield); S -

suitable (60–80%); MS - moderately suitable (40–60%); mS - marginally suitable (20–40%); NS - not suitable ($<$ 20% of maximum attainable yield).

For each crop type and grid-cell the starting and ending dates of the crop growth cycle are determined individually to obtain best possible crop yields, calculated separately for rain-fed and irrigated conditions. This procedure also guarantees maximum adaptation in simulations with year-by-year historical weather conditions, or under climate distortions applied in accordance with various climate change scenarios.

Adequate agricultural exploitation of the climatic potentials or maintenance of productivity largely depends on soil fertility and the use and management of the soil on an ecologically sustained basis. The AEZ agro-edaphic suitability rating scheme has been intensively used by the FAO and other organizations, at various scales, and in numerous countries and regions; it passed through several international expert consultations, and hence it constitutes the most recent consolidation of expert knowledge. In this system, suitability classifications are proposed for each soil unit, by individual crops at defined levels of inputs and management circumstances.

The model systematically tests the growth requirements of about 150 crop types against a detailed set of agro-climatic and soil conditions. For China the model operates on a 5 by 5 kilometer (km) grid; so the total grid matrix has 810 by 970 cells, of which about 375,000 grid-cells cover the mainland of China. Results of crop suitability analysis have been summarized in tabular and map form. On these maps, suitability results of each 5 km grid-cell of the China resource inventory are represented by a suitability index SI, reflecting the level of suitability of the part of each grid-cell considered suitable, and by the percentage suitable in that particular grid-cell.[1] The results for China, assuming an intermediate level of management and input conditions, are shown in *Figure 16.2*.

AEZ modeling establishes a platform for developing various applications. In LUC's study of China, besides providing land availability scenarios and the representation of agro-climatic and biophysical conditions for production function estimation, other recent examples deal with China's food prospects with special reference to water resources (Heilig *et al*., 2000). The AEZ modeling clearly brings out and quantifies the vast regional differences of China's rain-fed and irrigated grain production potential and irrigation water requirements.

The AEZ methodology is also ideally suited to analyzing the impacts of global climate change, as it permits the study of effects on the local water balance and is capable of integrating the impacts over a wide geographical area. Since all cropping possibilities (LUTs) are always tested, the AEZ method accounts for substitution where beneficial. By attempting to construct multiple crop combinations where the length of the growing season permits, the algorithm is fully adaptive to changing

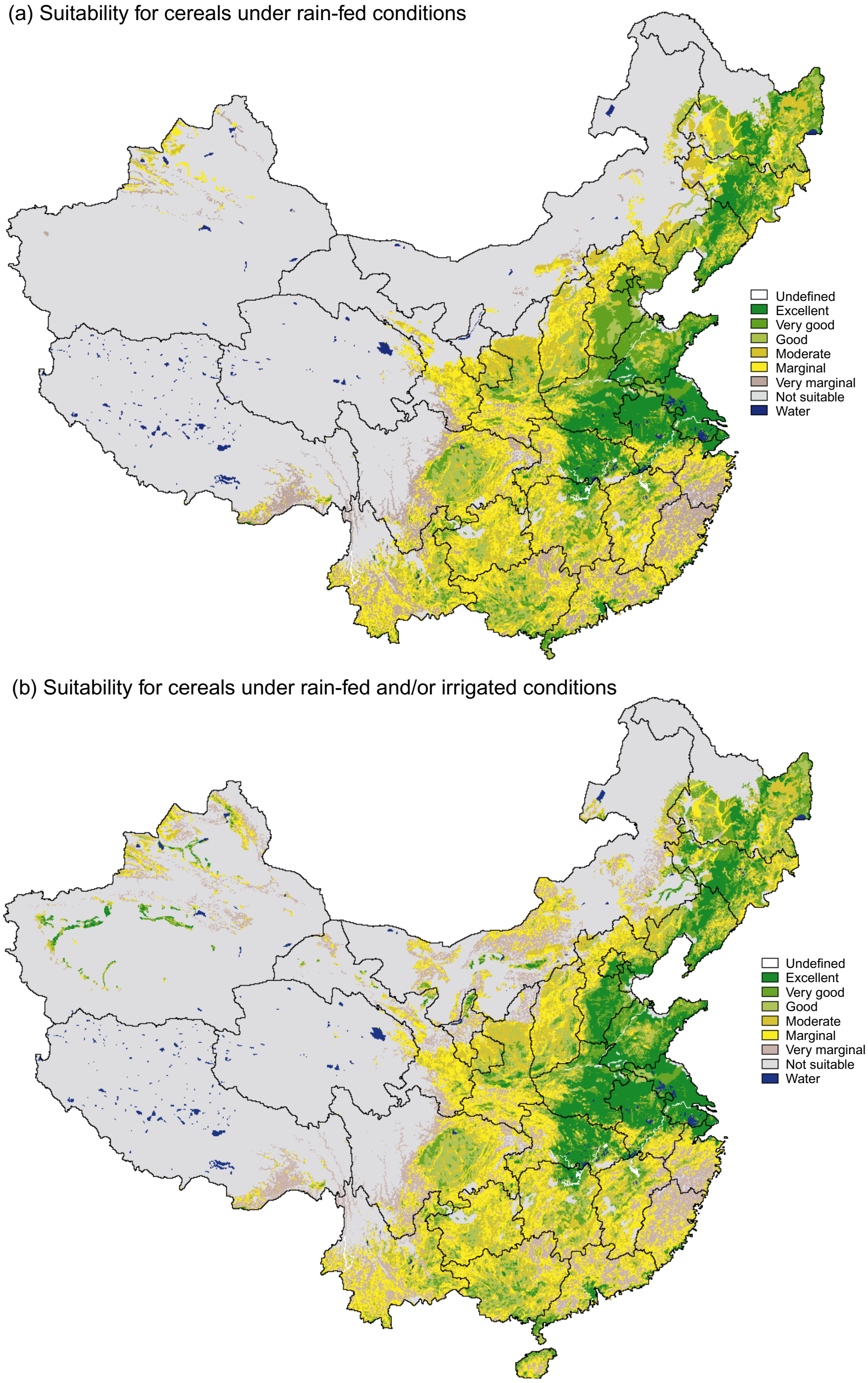


Figure 16.2. AEZ index SI of land suitability for cereal production. (a) Suitability for cereals under rain-fed conditions and (b) suitability for cereals under rain-fed and/or irrigated conditions.

conditions. Finally, AEZ assesses the crop production potential both under rain-fed and irrigated conditions and provides estimates of irrigation water requirements.

Before discussing climate change impacts, let us briefly review some results obtained for the reference climate conditions. Both temperature and rainfall distribution in China show very marked geographical patterns. For illustration, *Figure 16.3* shows the number of days with average daily temperatures above 5°C, the so-called thermal length of growing period LGPt. *Figure 16.4* presents annual rainfall, indicating a very pronounced southeast to northwest gradient ranging from about 2,000 millimeters (mm) per year in Guangdong province to less than 100 mm per year in parts of the Tibetian plateau and Xinjiang autonomous region. The distribution of cultivated land is shown in *Figure 16.5*.

The implication of China's agro-environmental conditions is that large parts of the North are water-stressed or water-deficient. This can be seen in *Figure 16.6*, which indicates the additional water required if the land were to be used for crop production during the entire time of the year when temperature would permit. The values shown represent additional (unmet) crop water requirements and do not include losses that would be incurred by providing water for irrigation. Assuming that irrigation efficiency may be in the order of 40% (FAO, 2000, Table 4.10, pp. 113), this means that the respective water supply would have to be 2.5 times the magnitude of crop water requirements.

Since water resources are best discussed in terms of river basins, we show a sketch of broad river systems (*Figure 16.7*). Since much of the cultivated land requiring irrigation for full exploitation of the production potential falls into the Huai He basin (number 3), the Huang He basin (number 4), and the Hei He-Luan He basin (number 2), our analysis of climate impacts on crop production potentials and water resources is concentrated on these three river basins.

Table 16.1 summarizes the importance of these water-stressed regions with regard to population distribution, food production, and cereal production potentials as calculated by AEZ.

As shown in *Table 16.1*, the three water-critical basins account for about 15 percent of China's land area, were home to 35% of China's population in 1990, and contributed also some 35% to China's food harvest in 1990. The rain-fed cereal production potential accounts for nearly 32% of the country's total whereas 34% would be contributed when considering rain-fed and/or irrigation conditions.

16.1.3 Water scarcity in the northeastern basins

Table 16.2 lists the values of common water stress indicators for the nine major watershed regions in China. The first index shows the per capita water availability. More than 2,000 cubic meters (m^3) of water per capita per year is considered sufficient water; 1,000 to 2,000 m^3 indicates a region under water stress, while less

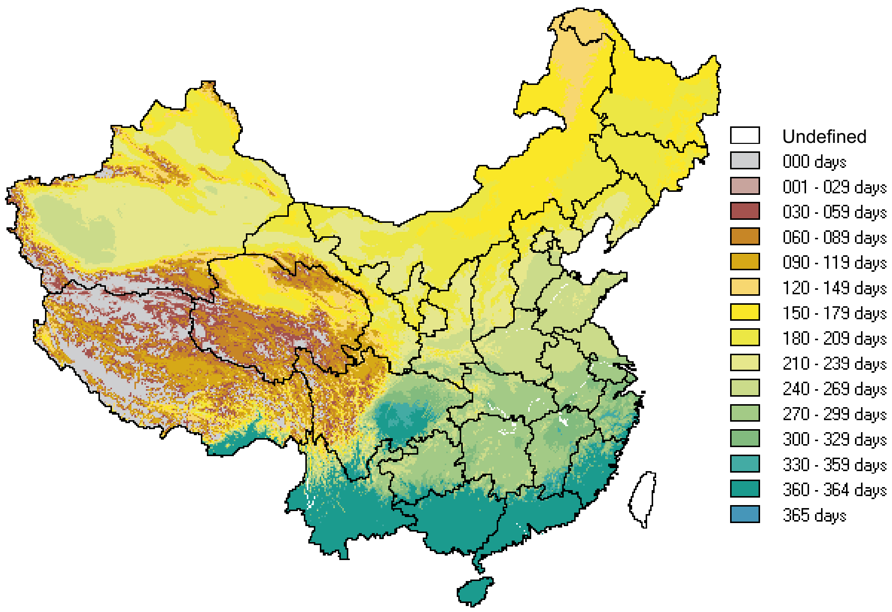


Figure 16.3. Number of days with average daily temperature >5°C.

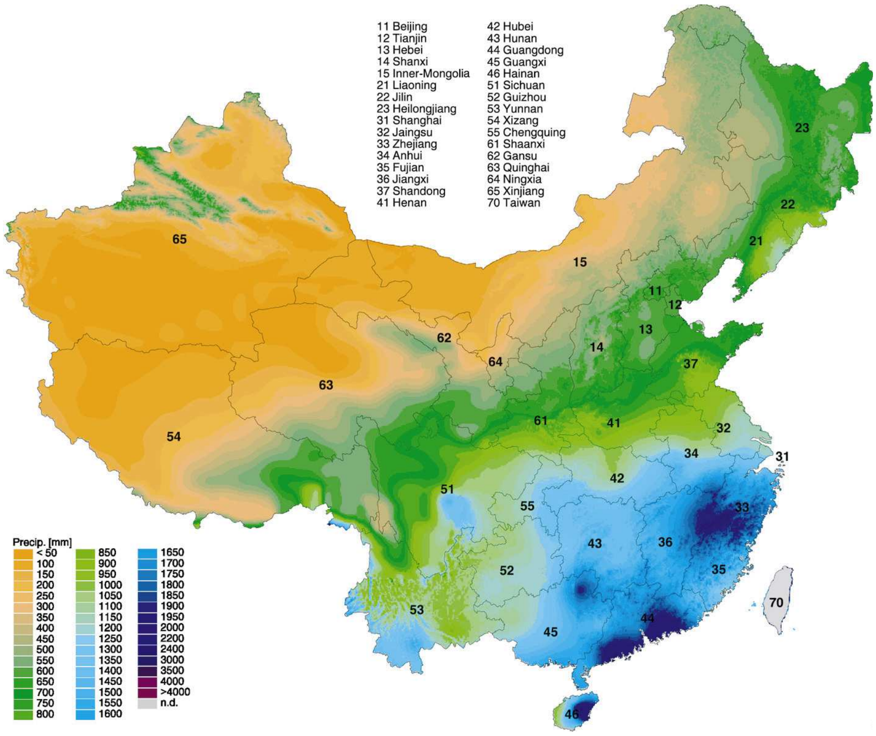


Figure 16.4. Mean annual precipitation (mm).

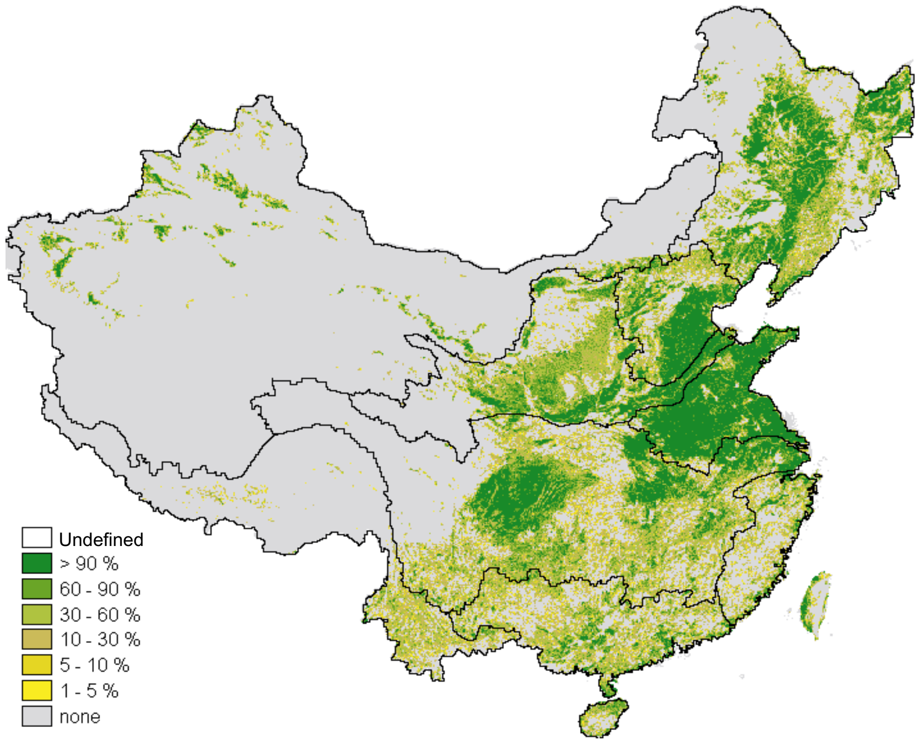


Figure 16.5. Distribution of cultivated land (percent of 5 km grid-cell).

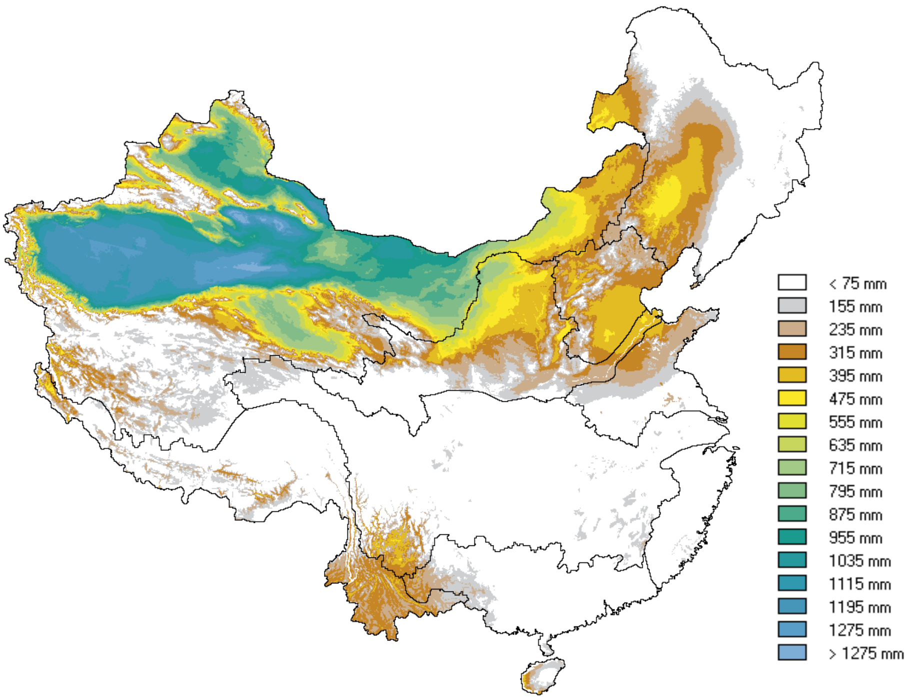


Figure 16.6. Crop water deficits assuming full exploitation of thermal growing season (mm).

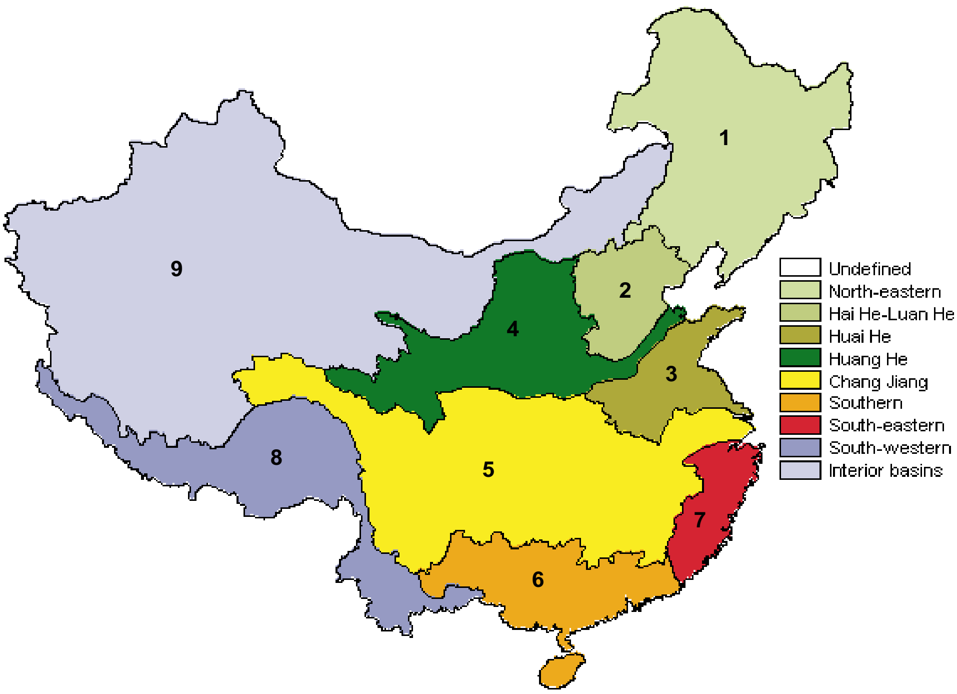


Figure 16.7. Major river basins in China.

Table 16.1. Selected indicators by water region in 1990 (percent in national total)

NR	River basin	Land area	Popu-lation	Food produc-tion	Culti-vated land	Irri-gated land	Rain-fed cereal potential	Rf + Ir cereal potential
I	Northeastern	13.2	9.8	14.7	18.2	7.3	17.8	17.9
II	Hai He-Luan He	3.4	9.2	8.1	9.5	10.4	7.6	8.9
III	Huai He	3.4	16.3	18.5	13.9	17.9	16.5	16.7
IV	Huang He	8.5	9.6	8.2	14.6	10.5	7.7	8.3
V	Chang Jiang	18.8	34.0	34.1	25.2	29.4	36.5	31.6
VI	Southern	5.8	11.3	8.6	9.0	10.6	10.0	9.2
VII	Southeastern	2.4	6.4	5.1	3.1	5.3	1.6	1.5
VIII	Southwestern	8.9	1.4	1.1	2.0	1.5	2.2	1.8
IX	Interior	35.5	1.9	1.7	4.5	7.1	0.1	4.1
	WR II + III + IV	15.3	35.1	34.8	38.0	38.9	31.8	33.9

than 1,000 m^3 indicates water scarcity.[2] Using this classification system, all four watershed regions in Northeastern China area water-stressed. Three of the four regions, the same regions which contain the NCP and close to 40% of the agricultural land, are severely stressed in the water scarcity category.

Table 16.2. Values of common water stress indexes by water region.

NR	River basin	Per capita surface water (m^3)	Demand/supply ratio 1993	2000	2010	Coefficient of variation	Storage/runoff	Security index	Combined index 1993	2000	2010
I	Northeastern	1,473	0.32	0.40	0.53	0.20	0.32	3	6	6	7
II	Hai He-Luan He	264	1.68	1.82	2.07	0.43	0.80	4	9	9	9
III	Huai He	395	0.97	1.15	1.39	0.33	0.16	5	10	10	10
IV	Huang He	551	0.74	0.82	1.04	0.31	0.68	3	8	8	8
V	Chang Jiang	2,403	0.21	0.24	0.28	0.12	0.18	4	6	6	6
VI	Southern	3,358	0.18	0.21	0.27	0.20	0.16	4	5	6	6
VII	Southeastern	3,346	0.13	0.15	0.19	0.21	0.15	5	6	6	6
VIII	Southwestern	35,850	0.01	0.02	0.02	0.08	0.03	3	3	3	3
IX	Interior	5,172	0.55	0.60	0.69	0.10	0.38	2	5	5	5
	China Total	2,306	0.23	0.26	0.31	0.07	0.17	3	5	5	5

The water stress in these regions is reinforced by the other indexes also. In 1993, the Huang He Basin had the lowest demand to supply ratio at 0.74.[3] In the other two water scarce basins, demand already equaled or exceeded surface water supply. In the Hai He-Luan He Basin, demand exceeded supply by close to 70%! According to demand projections by the Nanjing Institute of Hydrology and Water Resources, by 2010 demand will exceed surface water supply in all three regions. Currently, groundwater pumping supplies the excess demand. However, in many areas this practice is not sustainable, with water tables dropping up to 5 m/year at some wells, resulting in land subsidence and saltwater intrusion in coastal areas.

Adding to the difficulty of water-stress in these regions is great inter- and intra-annual variability in runoff. The inter-annual coefficient of variation is above 0.3 in all three regions, indicating both floods and droughts are common. Intra-annual variation in these regions is also extreme. Here the coefficient of variation is greater than 1 in all cases, with 75% of the runoff occurring between July and October. Building storage reservoirs can, however, effectively reduce the variation in flow. In the Hai He–Luan He Basin, water resources availability has become more dependable with storage capacity in the basin at more than 80% of annual average runoff. The Huang He Basin is also well developed. However, the Huai He Basin does not have as many good dam sites and cannot develop the level of storage seen in the other watersheds. This fact alone makes the combined stress index for this basin the worst of the three, although the average annual water available is greater than in the Hai He–Luan He basin.

Figure 16.8 illustrates the degree of current and expected future water stress in the Hai He-Luan He, Huai He, and Huang He basins by showing current surface water supply along with projected demand estimates for 1993, 2000, and 2010. The

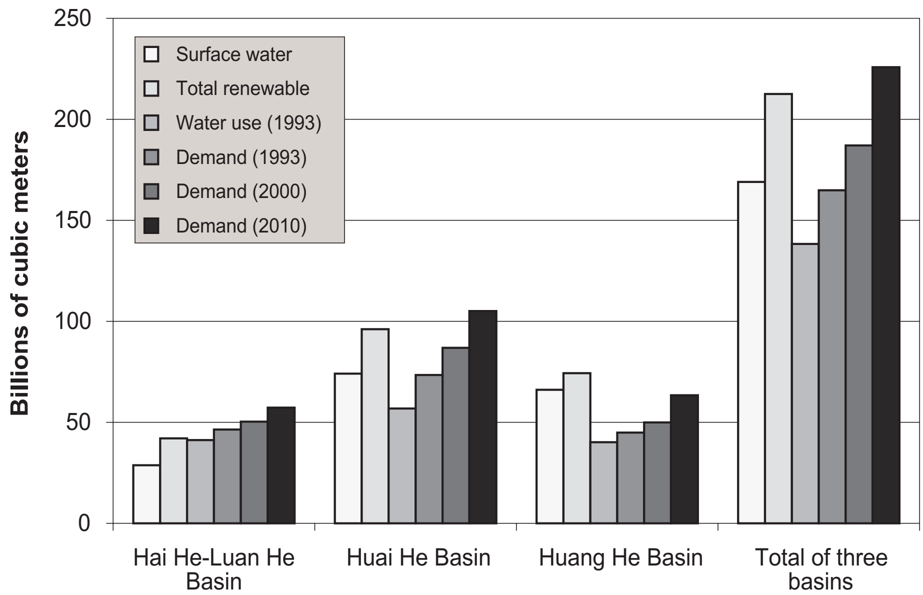


Figure 16.8. Comparison of water supply, use, and three demand scenarios. Source: Report on the mid- and long-term plans for water demand and supply in China, Nanjing Institute of Hydrology and Water Resources, November 1996.

figure also indicates the estimated renewable water supply calculated by adding a measure of renewable groundwater supplies to the surface water runoff (UN, 1997). The plot reinforces the results of *Table 16.2*, indicating that by 2010 demand in the three regions will outstrip not only surface water supply but even the estimated total renewable water supply in the regions. As previously stated, groundwater is already being extracted at a rate that exceeds the renewable supply. If groundwater cannot even continue to be exploited at current levels, water must be brought in from other, wetter regions of China to satisfy demand.

Improvements in water supply systems and irrigation efficiency have a great potential to reduce irrigation water demand. However, these improvements come at the cost of reduced groundwater recharge, resulting in perhaps less groundwater availability. In addition, growth in domestic and industrial demand for water dwarf the growth in agricultural demand as shown in *Figure 16.9*. Agriculture is expected to receive a diminishing percentage of the water in these regions as cities require much more. Urban growth in China has been so fast that the urban supply capacity has increased 4.6 times from 1978 to 1993 and is still unable to meet demand (UN, 1997, p. 22). Considering the severe water constraints in these populous agricultural and industrial regions, assessing the impacts of climate change on water resources in this area becomes very important.

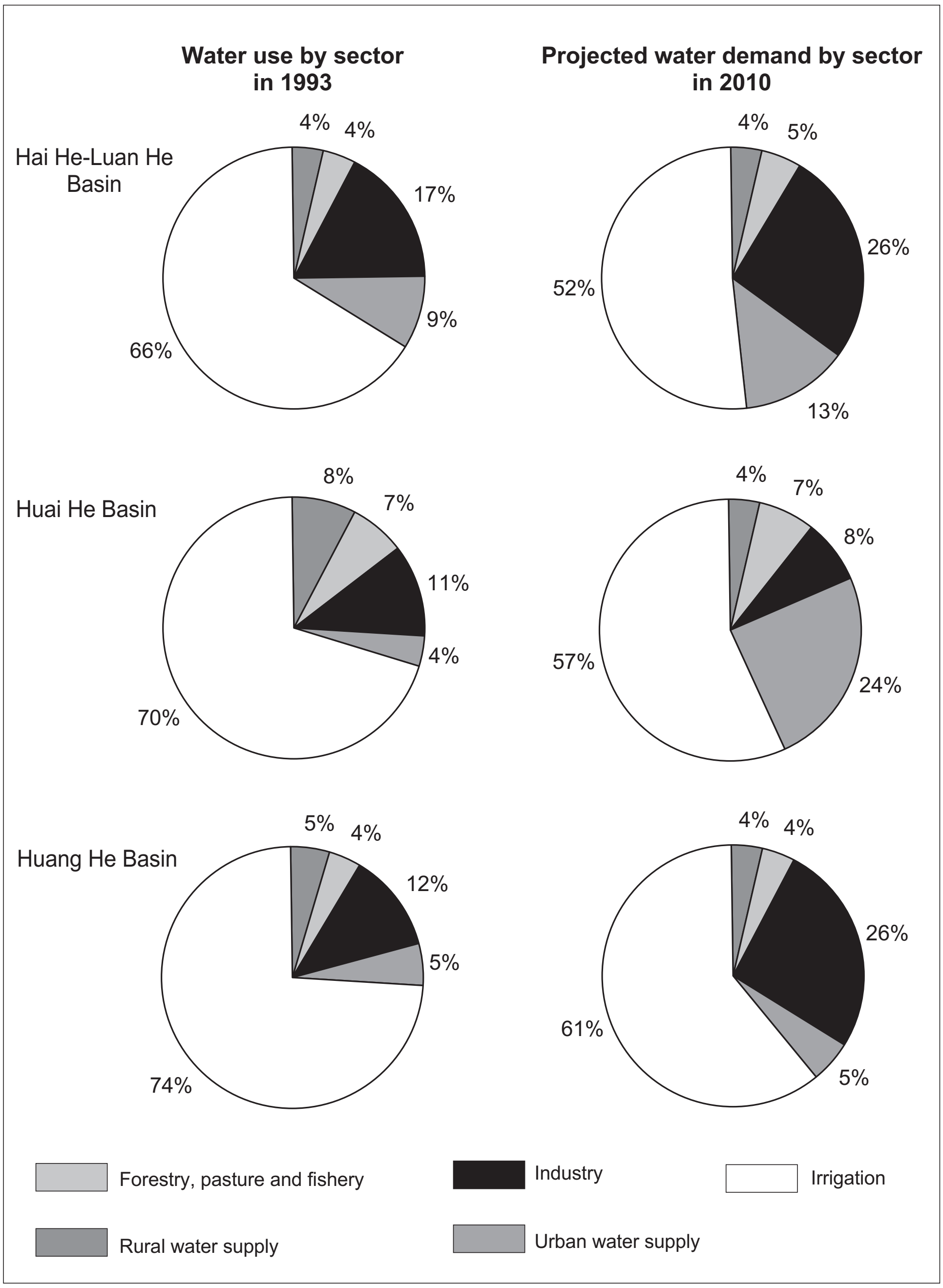


Figure 16.9. Current and estimated future water demand by sector.

16.2 Modeling Climate Change Impacts on Land and Water Resources in North China

A number of general circulation model (GCM)-based climate scenarios were selected for use in AEZ land productivity assessments. Outputs from six GCM experiments were obtained through the Intergovernmental Panel on Climate Change's Data Distribution Centre (DDC). They include the following models/scenarios for the periods 2010–2039, 2040–2069, and 2070–2099:

1. *The ECHAM4 model*. This model was developed at the German Climate Research Center of the Max-Planck-Institute for Meteorology in Hamburg, Germany (Oberhuber, 1993; Roeckner *et al*., 1992; Roeckner *et al*., 1996). Model results were taken from the greenhouse-gases-forcing scenario and from the greenhouse-gases-plus-sulfate-aerosols-forcing scenario. For the latter only the 2010–2039 period was available. The scenario results from ECHAM4 are provided at spatial resolution of approximately 2.8 $\times$ 2.8 degrees.
2. *The Canadian Global Coupled Model (CGCM1)*. This model was developed at the Canadian Centre of Climate Modelling and Analysis. Model results were taken from the greenhouse-gases-forcing scenario and from the average of "ensemble" simulations (ensemble simulations are based on identical historical and future changes in greenhouse gases, however initiated from different points on the control run). The average "ensemble-forcing scenario" was taken for the greenhouse-gases-plus-sulfate aerosols. The scenario results from CGCM1 are provided at spatial resolutions of 3.75 $\times$ 3.75 degrees (Boer *et al*., 2000; Flato *et al*., 2000).
3. *The HadCM2 model*. This model is based on recent experiments performed at the Hadley Center for Climate Prediction and Research (Murphy, 1995; Murphy and Mitchell, 1995). Model results were taken from the average of "ensemble" simulations. The average of "ensemble-forcing scenarios" was used for greenhouse gases only and for greenhouse-gases-plus-sulfate aerosols, respectively. The scenario results from HadCM2 are available at a spatial resolution of 3.75 $\times$ 2.75 degrees.

Outputs of the above six climate model experiments, available for three time periods and with various spatial resolutions, have been interpolated to 0.5 $\times$ 0.5 degrees.

At a minimum, four climatic parameters from the GCM results were used to adjust the baseline climate conditions of each grid-cell. The *difference* in monthly mean maximum and minimum temperatures, between a GCM climate change run and the respective GCM control experiment (representing approximately the cur-

Table 16.3. Scenarios of global climate change.

Period/Scenario	ECHAM4	HadCM2	CGCM1
North China		Mean temperature change (°C)	
2020s	1.5	1.6	3.1
2050s	3.0	2.6	5.7
2080s	4.4	3.6	10.0
North China		Annual precipitation (% change)	
2020s	8.7	10.4	–1.7
2050s	16.4	11.3	9.4
2080s	20.2	30.5	–11.1

rent base climate), was added respectively to the mean monthly maximum and minimum temperatures of the baseline climate surfaces. Multipliers, i.e., the *ratio* between the GCM climate change and control experiment, were used to impose changes in precipitation and incident solar radiation, respectively. When available from a GCM, changes in wind speed and relative humidity were considered as well. Each climate scenario is also characterized by the level of atmospheric CO_2 concentrations and assumed changes in crop water-use efficiency. These parameters affect both the estimated reference evapotranspiration as well as the crop biomass estimations.

The average annual temperature and precipitation changes projected for the NCP are shown in *Table 16.3*. Changes in precipitation derived from the various GCM results are displayed in *Figure 16.10*, showing the seasonal distribution for 3-month periods starting with winter (DJF = December, January, February) through autumn (SON = September, October, November). Note that there are some clear differences in the seasonal projections, which we will again refer to when discussing consequences for surface water resources.

16.3 Climate Change Impacts on Crop Production Potentials

From the current distribution of rainfall dominated by the summer monsoon and the seasonal temperate and sub-tropical climate in most of China, it can be concluded that temperature increases could be beneficial as long as the increased evapotranspiration is compensated for by additional rainfall. Also, a change of rainfall distribution, shifting precipitation towards the water-deficit months in spring, will generally improve the conditions for rain-fed cropping. Hence, for the NCP, the calculated climate change impacts on cropping and water resources are governed by a subtle combination of a reduction of temperature constraints to cropping and

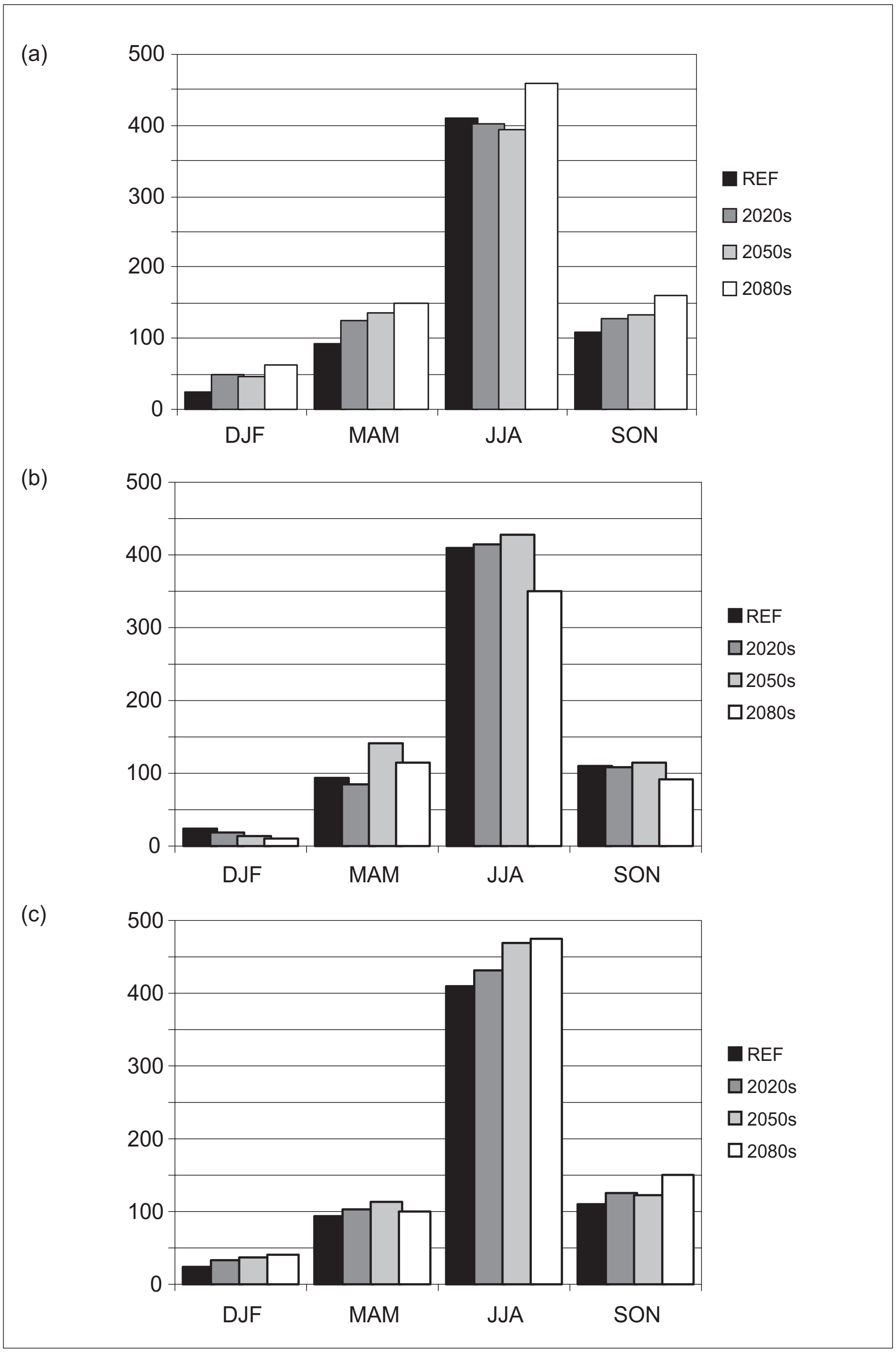


Figure 16.10. Current and future seasonal precipitation in North China. (a) Hadley Center Model, HadCM2 (ensemble forcing); (b) Canadian Global Coupled Model, CGCM1; and (c) German Climate Research Center Model, ECHAM4.

Table 16.4. Climate change impacts on cereal production potential in North China (percent change with respect to reference climate).

	HadCM2		CGCM1		ECHAM4	
	2050s	2080s	2050s	2080s	2050s	2080s
Rain-fed						
Area (VS+S+MS)	0.2	6.3	2.4	–18.5	15.0	20.3
Average yield / ha	9.2	12.8	–12.4	–14.1	–3.2	–8.1
Potential production	9.5	19.9	–10.3	–30.0	11.3	10.6
Suitability index	3.2	7.7	–1.1	–19.1	13.7	16.4
Rain-fed and irrigated						
Area (VS+S+MS)	1.4	7.3	7.4	–8.2	14.7	19.9
Average yield / ha	8.3	13.0	3.8	13.1	–0.5	–1.3
Potential production	9.8	21.3	11.5	3.8	14.1	18.3
Suitability index	2.4	5.0	–0.5	–12.3	10.0	12.4
Irrigation water	3.1	–29.9	30.9	150.8	–4.6	–2.9

changes, positive or negative depending on respective GCM results, in the water balance.

Table 16.4 presents results for cereal production potentials under rain-fed as well as rain-fed + irrigated conditions. The calculations assume that all adaptations a rational farmer would be able to undertake are actually taken. In other words, the computations determine for each grid-cell the crop or crop combination that would maximize cereal output from all suitable areas in that grid-box. Hence, shifting of crop calendars, change of crop type, change of crop, and change of multi-cropping are all considered, separately for rain-fed and irrigation conditions.

When considering the future climate based on the results of the Hadley Center GCM (HadCM2; average of ensemble-forcing scenarios for greenhouse gases only), conditions for agriculture in the three river basins encompassing most of the NCP improve significantly, both for rain-fed cropping and for irrigated agriculture (see first two columns with results in *Table 16.4*). In order not to confound the thermal and water balance changes with (somewhat uncertain) impacts due to CO_2 fertilization, the latter have been excluded from *Table 16.4*. Even then, the production potentials for rain-fed and rain-fed + irrigated cereal increase around 10% for the 2050s, and by some 20% for the 2080s. As annual precipitation is projected by the HadCM2 model to increase by 30%, this more than compensates for

Table 16.5. Climate change impacts on cereal production potential in mainland China (percent change with respect to reference climate).

	HadCM2		CGCM1		ECHAM4	
	2050s	2080s	2050s	2080s	2050s	2080s
Rain-fed						
Area (VS+S+MS)	3.4	6.8	5.7	2.7	13.0	19.0
Average yield / ha	0.4	3.5	–8.1	–7.0	–6.3	–7.9
Potential production	3.8	10.6	–2.9	–4.5	5.8	9.6
Suitability index	6.2	8.6	3.8	–0.2	13.2	16.4
Rain-fed and irrigated						
Area (VS+S+MS)	4.7	7.8	8.6	7.6	13.4	18.4
Average yield / ha	1.9	5.5	0.7	4.6	–3.3	–2.4
Potential production	6.7	13.8	9.3	12.6	9.6	15.6
Suitability index	6.8	8.5	5.2	2.8	12.3	14.8
Irrigation water	12.3	–4.2	42.6	114.6	11.4	21.3

a temperature rise of 3.6°C. Consequently, the crop water balance improves (as is apparent from the improvements in the Suitability Index) and less irrigation water is required to fully exploit the cereal production potential. For comparison, we also present the aggregate results for all of mainland China (see *Table 16.5*).

Impacts using the results of the Canadian climate model (CGCM1) are quite opposite. The model produces a large warming of 5.7°C and 10.0°C for the 2050s and 2080s, respectively. Precipitation increases by almost 10% for the 2050s, but then decreases by more than 10% below the reference climate levels. This translates into a large reduction of rain-fed cereal production potentials. With irrigation, this trend could be reversed. However, irrigation requirements for full exploitation of the thermal resources would result in a 30% irrigation water increase in the 2050s, and a 2.5-fold increase for the 2080s. Obviously, such vast amounts of water would hardly be available, even when considering South-North water transfers.

Finally, the ECHAM4 model developed by the Max-Planck Institute projects a climate change comparable in magnitude to that of the Hadley Center. Temperature rises 3.0°C and 4.4°C, respectively, for the 2050s and 2080s; precipitation increases by 16 and 20%, respectively. As a consequence, rain-fed and rain-fed + irrigated cereal production potentials increase by about 10% for both time periods. The future climate based on ECHAM4 results produces the largest increases in suitable areas, some 20% for the 2080s. However, this large expansion of cropping into currently dry and less productive areas results in some reduction of average output per hectare of suitable area.

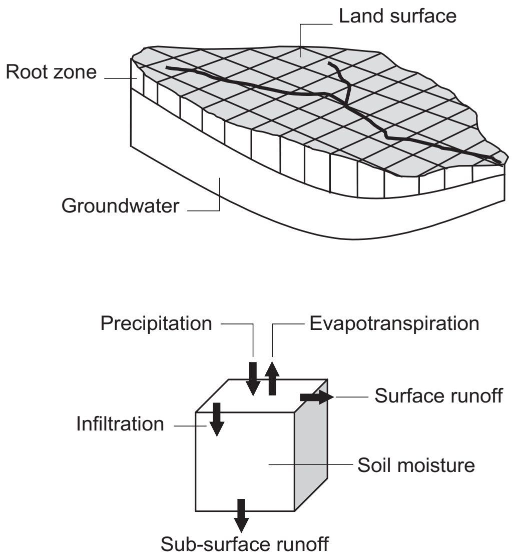


Figure 16.11. Illustration of CHARM surface runoff model.

16.4 Modeling Climate Change Impacts on Surface Water Resources

A model called CHARM (**C**limate and **H**uman-**A**ctivities sensitive **R**unoff **M**odel) developed by the LUC project at IIASA was used to model the impact of climate change on runoff in the Hai He–Luan He, Huai He, and Huang He basins. As depicted in *Figure 16.11*, CHARM first calculates direct runoff and then performs a soil water balance at a grid-cell level. Sub-surface runoff that later enters back into streams is lumped together for the entire basin and calibrated with a single parameter (Wiberg and Strzepek, 2000). A base scenario was obtained by simulating sixteen years of data from 1965 to 1980. The use of time series data allows for calculating variability of runoff as well as averages.

Figure 16.12 shows the results of applying CHARM to the changed climate of the Hadley Center, Canadian, and Max-Planck climate scenarios for the 2050 decade. The figure shows that although rainfall is increased in all three scenarios, runoff decreases in the HadCM2 scenario while the other two GCM scenarios result in an increase in runoff. The slightly decreasing runoff in the HadCM2 scenario is a result of the timing of rainfall throughout the year. The Hadley scenario provides more even rainfall throughout the year, increasing the rainfall in dry months and decreasing the rainfall in the wettest months. This was indicated previously in *Figure 16.10* and again here in *Figure 16.13*, which shows the distribution of runoff throughout the year. During the dry months, the soil is quite capable of

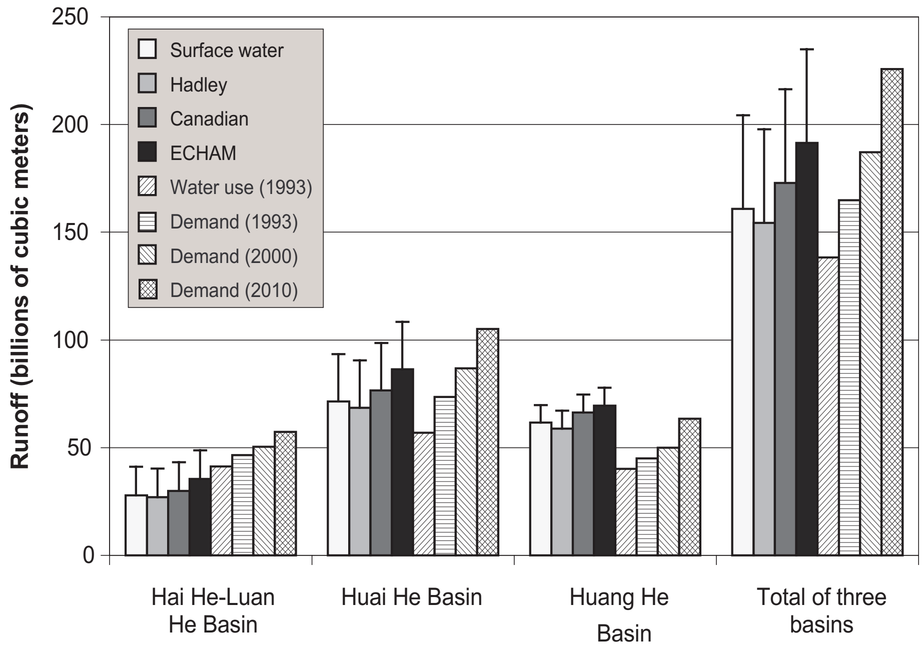


Figure 16.12. Impact of climate change scenarios (in 2050s) on surface water runoff. Lines above bars represent addition of estimated sustainable groundwater withdrawals.

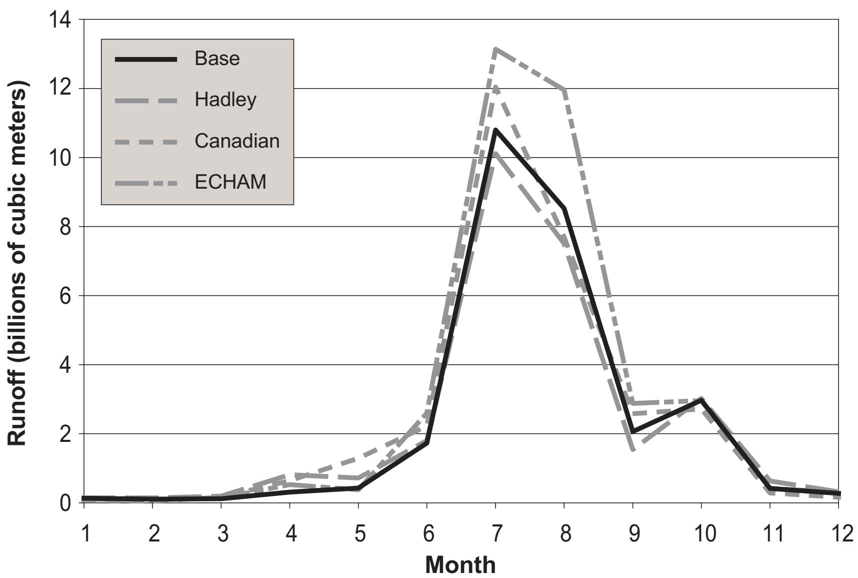


Figure 16.13. Monthly surface water runoff in various climate change scenarios (in 2050s).

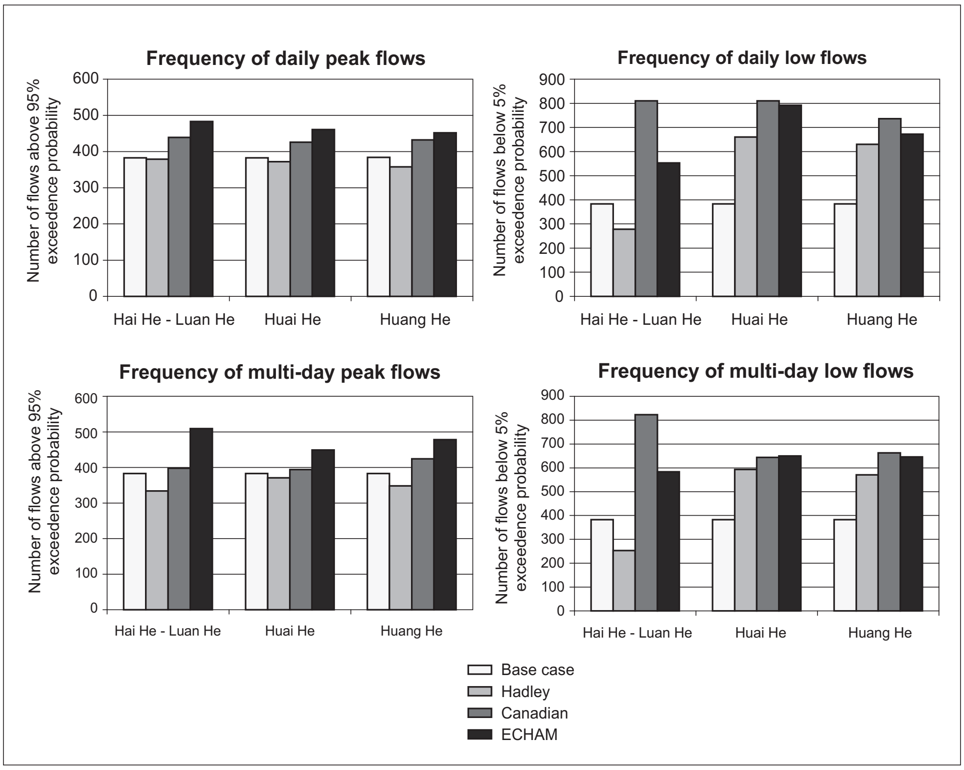


Figure 16.14. Frequency of extreme events in river basins of North China (days during 16 years of simulation, present, and in 2050s).

absorbing and evaporating more water. Although rainfall increases substantially in these months, runoff may not increase much due to storage in soils. In the wet months, the soil is practically always saturated, so that any additional rainfall simply runs off. However, the HadCM2 scenario reduces rainfall in these months, which consequently reduces the runoff for the year. Both the CGCM1 and ECHAM models increase rainfall in months that already have the greatest rainfall, producing much greater runoff in these months.

Even with the most favorable conditions in the ECHAM scenario, however, projected water demand in 2010 will already outstrip surface water supply (of 2050s) in the three regions. If the estimated sustainable groundwater withdrawals are added (as shown by the small lines above the bars in *Figure 16.12*), total water supply in the three regions can finally about meet 2010 demand, but 2050 demand is likely to be higher. Furthermore, the safe yield from aquifers could change with management and climate change. Demand cannot be met in the Hai He–Luan He basin in any scenario without importing water from other regions.

Variability of flows can be even more important than average flows when developing water resources. Higher variability means greater frequency and severity of floods and droughts and the associated damages. It also means that more storage capacity must be built to save water during periods of high flow for periods of drought. This greater storage capacity also helps to reduce the risk of flood and drought. *Figure 16.14* illustrates the impact of the climate scenarios on the frequency of extreme events. Two measures are used, daily flows and multi-day flows below the 5% and above the 95% probability flow level of the base scenario. In this case "multi-day" means a 14-day moving sum, which better represents a flood or drought period than a one-day flow. The frequency of peak flows follows the same pattern as the average runoff under the three scenarios. However, the change from the base case is slightly more pronounced. For instance, in the Canadian scenario, average runoff increases by 7%, but the number of peak flows increases by nearly double that percentage in all cases. The most extreme change, though, is the number of two-week periods with very low flow. These drought periods increase in all scenarios but one and increase quite substantially, between 50% and 100%.

Although the frequency of floods and droughts appears to increase under the climate scenarios, the prediction of more water and a slightly more even distribution of the water throughout the year in some scenarios results in a positive water yield effect at existing storage capacity. As shown in *Figure 16.15*, *Figure 16.16*, and *Figure 16.17*, the water yield from the current storage capacity actually increases in some cases. At least, the yield does not decrease under any of the scenarios. The limitation to this conclusion is that the curves illustrated do not account for evaporation. Evaporation is highly dependent on reservoir shape, which is unknown in future development and so was not included here. However, evaporation should increase in all of the future climate scenarios for this region, resulting in smaller water yields than shown here. Furthermore, at high reservoir capacity levels, as storage is further increased, the curves will actually show decreasing yields, since evaporation losses will increase quickly with greater reservoir surface area.

16.5 Conclusions

Climate change will affect China's agriculture in many ways. Results from three global climate modeling experiments were applied to two complementary environmental assessment models to quantify likely impacts of global climate change on multiple cropping systems, cereal production potentials, and surface water resources in North China. This important agricultural region is water-stressed under ambient conditions and is very sensitive in its soil moisture balance to temperature and precipitation changes.

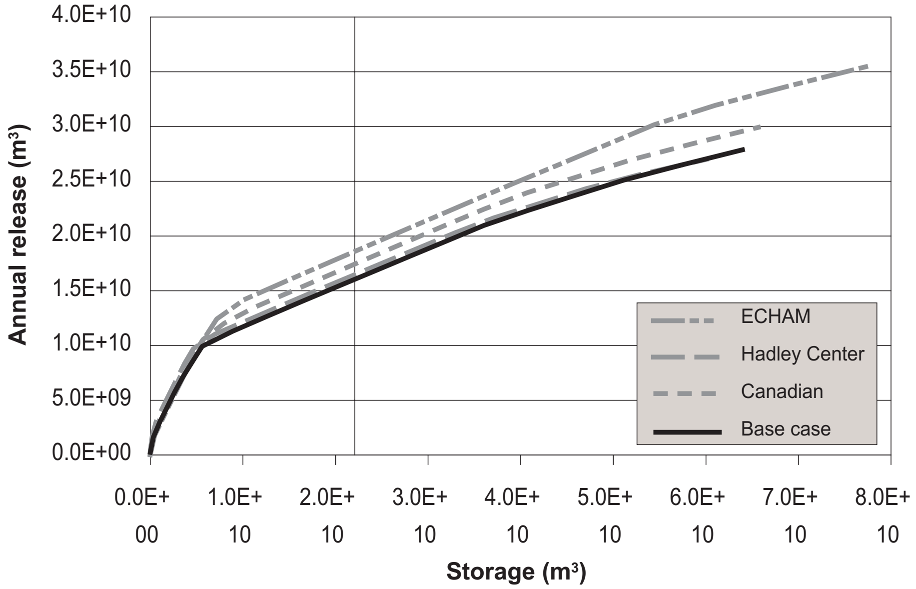


Figure 16.15. Water yield for Hai He-Luan He under various climate change scenarios (2050s).

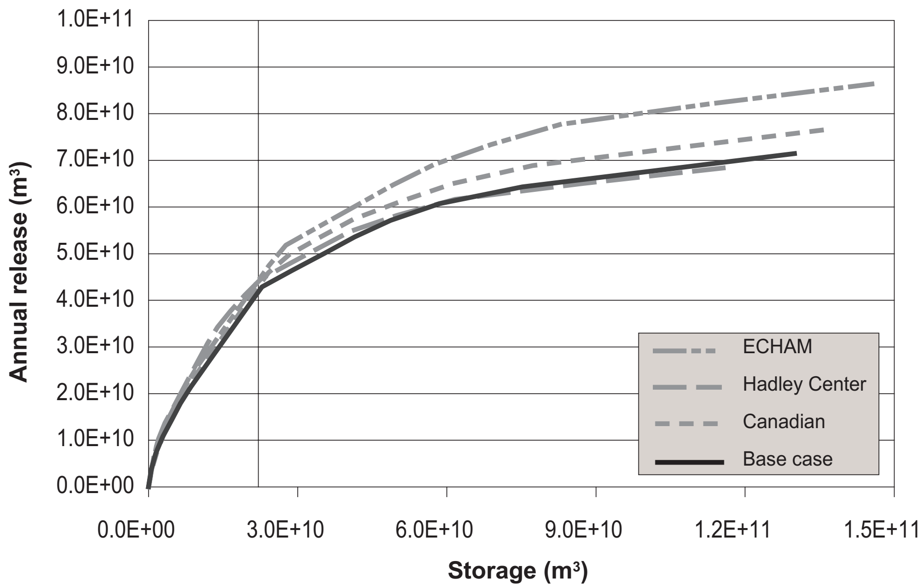


Figure 16.16. Water yield for Huai He basin under various climate change scenarios (in 2050s).

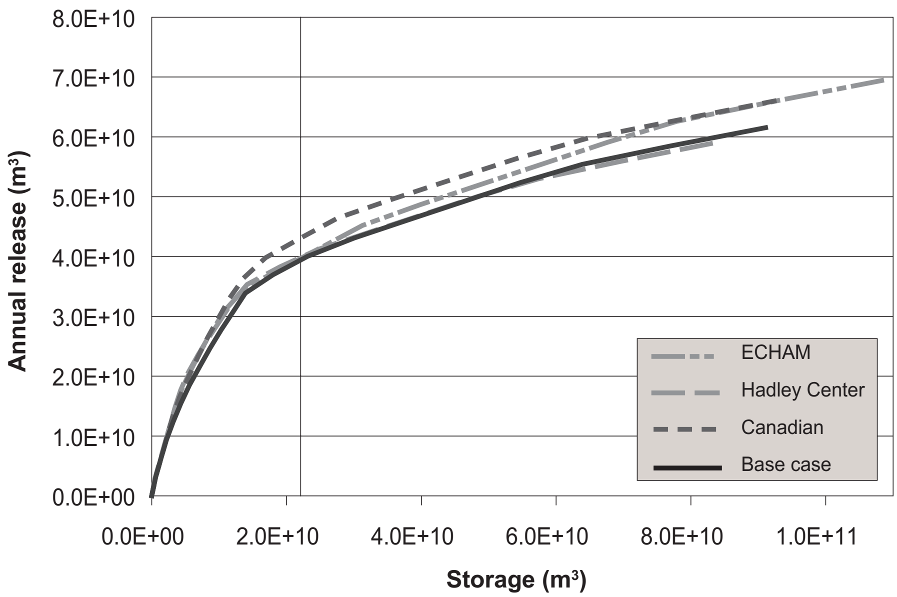


Figure 16.17. Water yield for Huang He basin under various climate change scenarios (in 2050s).

Temperature increases as projected for the NCP by the Hadley Center (HadCM2) and Max-Planck Institute (ECHAM4) coupled ocean-atmosphere climate models range from 1.5–1.6°C in the 2020s, 2.6–3.0°C in the 2050s, and 3.6–4.4°C in the 2080s. The Canadian model predicts much more dramatic increases, about twice as high. Rainfall increases quite strongly for ECHAM4 and HadCM2 models. The Canadian model (CGCM1) produces irregular and changing responses for different time periods, most negative for the 2080s.

With its seasonal temperate climate, much of North China's agricultural capacity is both temperature- and water-limited. Global warming will alleviate temperature constraints and harsh winter conditions and will increase opportunities for multi-cropping, provided water is available. With the projected changes in agro-environment, rain-fed cereal production potentials in the NCP respond positively, with increases of 10–20%, in the case of ECHAM4 and HadCM2. The severe drought conditions projected with the Canadian climate model cause a decline of rain-fed cereal production potential by 30%. If water were available for irrigation, the rain-fed + irrigated production potential would exceed the levels of the reference case for all climate runs. However, especially for conditions projected with the Canadian CGCM1 climate model, very massive irrigation expansion would be required, which is quite unlikely under warmer and dryer conditions in the three river basins of the NCP.

Surface water runoff depends on the total amount of precipitation as well as its seasonal distribution. Changes in the latter may cause a reduction of surface water runoff even when total precipitation increases, such as occurs in HadCM2 results for the 2050s. In most cases, however, the surface water runoff in the NCP improves with climate change. Even then, the model results suggest increased occurrence of extreme flow events.

In summary, the AEZ and CHARM model simulations based on three GCM experiments suggest that water will become an increasingly decisive factor in northern China. Even without climate change, competition for water is expected to grow due to projected population increases and urban/industrial growth. Warming results in an increasing crop water demand and will render the water balance in the NCP more vulnerable to rainfall variability, especially droughts. While two of the climate models project conditions that would lead to an overall improvement, both in the NCP and in mainland China, one climate model predicts more severe drought conditions with rather negative impacts on agriculture. Hence, given the wide range of climate predictions, it is not possible to come up with a firm and narrow answer as to what global environmental change will mean for China in the long-term. What is possible, however, is to confirm that any achievements in agricultural technology, improvements in irrigation water use efficiency, and incentives to economic use of water will be beneficial now and perhaps absolutely necessary in the future.

Notes

[1] The suitability index is defined as SI = 0.9 VS + 0.7 S + 0.5 MS + 0.3 mS.

[2] Sandra Postel uses this as a scarcity index in her book *Last Oasis – Facing Water Scarcity*. She points to Malin Falkenmark as being responsible for the definition and refers to one of her papers, "The Massive Water Scarcity Now Threatening Africa - Why Isn't it Being Addressed?" *Ambio*, Vol. 20, No. 1, 1991. Shiklomanov (1993, 2000) arrives at a similar scarcity index by subtracting unrecoverable water consumption from total runoff and dividing by population. In Shiklomanov's grouping, $<$1,000 cubic meters per capita per year is considered catastrophically low, 1,000–2,000 is very low, 2,000–5000 is low, 5,000–10,000 is average, 10,000–20,000 is high, and $>$20,000 is very high.

[3] Falkenmark and Lindh (1993) state that "Many countries, therefore, consider 30–60% of theoretically available water resources to be the practical limit of what they can mobilize." They go on to say that 20% may be a better estimate in the short to medium term for developing countries, since costs of water development have become " . . . increasingly dominant in national economies" in the developed countries that have gone above this point.

References

Boer, G.J., Flato, G.M., Reader, M.C., and Ramsden, D., 2000, A transient climate change simulation with historical and projected greenhouse gas and aerosol forcing: experimental design and comparison with the instrumental record for the 20th century, *Climate Dynamics*, **16**:405–425.

Falkenmark, M., and Lindh, G., 1993, Water and economic development, in P.H. Gleick, ed., *Water in Crisis: A Guide to the World's Fresh Water Resources*, Oxford University Press, Oxford, UK and New York, NY, USA.

FAO (Food and Agriculture Organization), 1984, Guidelines: Land evaluation for rainfed agriculture, FAO Soils Bulletin 52, Food and Agricultural Organization of the United Nations, Rome, Italy.

FAO (Food and Agriculture Organization), 1985, Guidelines: Land evaluation for irrigated agriculture, FAO Soils Bulletin 55, Food and Agricultural Organization of the United Nations, Rome, Italy.

FAO (Food and Agriculture Organization), 2000, Agriculture: Towards 2015/30, Technical Interim Report, April, Global Perspective Studies Unit, Food and Agricultural Organization of the United Nations, Rome, Italy.

Fischer, G., Frohberg, K., Parry, M.L., and Rosenzweig, C., 1996, Impacts of potential climate change on global and regional food production and vulnerability, in E.T. Downing, ed., *Climate Change and World Food Security*, Springer–Verlag, Berlin, Germany.

Fischer, G., forthcoming, Integrating biophysical and socioeconomic factors in modeling impacts of global environmental change, submitted for publication to the Terra Scientific Publishing Co., Tokyo, Japan, scheduled for publication June 2001.

Flato, G.M., Boer, G.J., Lee, W.G., McFarlane, N.A., Ramsden, D., Reader, M.C., and Weaver, A.J., 2000, The Canadian Centre for Climate Modelling and Analysis global coupled model and its climate, *Climate Dynamics*, **16**:451–467.

Heilig, G.K., Fischer, G., and Velthuizen, H. van, 2000, Can China feed itself? An analysis of China's food prospects with special reference to water resources, *Int. J. Sustain. Dev. World Ecol.*, **7**:153–172.

Murphy, J.M., 1995, Transient response of the Hadley Centre Coupled Model to increasing carbon dioxide. Part I: Control climate and flux adjustment, *J. Climate*, **8**:496–514.

Murphy, J.M., and Mitchell, J.F.B., 1995, Transient response of the Hadley Centre Coupled Model to increasing carbon dioxide. Part II: Temporal and spatial evolution of patterns, *J. Climate*, **8**:57–80.

Nanjing Institute of Hydrology and Water Resoures, 1996, Report on the mid- and long-term plans for water demand and supply in China, Nanjing Institute of Hydrology and Water Resources, PRC, Nanjing, China (in Chinese).

Oberhuber, J.M., 1993, Simulation of the Atlantic circulation with a coupled sea-ice mixed layer-isopycnal general circulation model. Part I: Model description, *J. Phys. Oceanogr.*, **13**:808–829.

Postel, S., 1992, *Last Oasis: Facing Water Scarcity*, W.W. Norton and Company, New York, NY, USA.

Roeckner, E., Arpe, K., Bengtsson, L., Brinkop, S., Dümenil, L., Esch, M., Kirk, E., Lunkeit, F., Ponater, M., Rockel, B., Suasen, R., Schlese, U., Schubert, S. and Windelband, M., 1992, Simulation of the present-day climate with the ECHAM 4 model: impact of model physics and resolution, Max-Planck Institute for Meteorology, Report No. 93, Hamburg, Germany.

Roeckner, E., Arpe, K., Bengtsson, L., Christoph, M., Claussen, M., Dümenil, L., Esch, M., Giorgetta, M., Schlese, U., and Schluzweida, U., 1996, The atmospheric general circulation model ECHAM-4: model description and simulation of present-day climate, Max-Planck Institute for Meteorology, Report No. 218, Hamburg, Germany.

Rosenzweig, C., and Parry, M.L., 1994, Potential impacts of climate change on world food supply, *Nature*, **367**:133–138.

Shiklomanov, I., 1993, World fresh water resources, in P.H. Gleick, ed., *Water in Crisis: A Guide to the World's Fresh Water Resources*, Oxford University Press, Oxford, UK and New York, NY, USA.

Shiklomanov, I., 2000, Chapter 12, World water resources and water use: Present assessment and outlook for 2025, in *World Water Scenarios: Analysis*, World Water Vision, http://www.watervision.org.

Turner II, B.L., Skole, D., Sanderson, S., Fischer, G., Fresco, L., and Leemand, R., 1995, Land-Use and Land-Cover Change: Science/Research Plan, IGBP Report No. 35 and HDP Report No. 7, IGBP, ICSU, HDP and ISSC, Stockholm, Sweden and Geneva, Switzerland.

UN (United Nations), 1997, Study on Assessment of Water Resources of Member Countries and Demand by User Sectors. China: Water Resources and Their Use, United Nations, New York, NY, USA.

UNDP/SSTC/FAO/SLA, 1994, Land Resources, Use and Productivity Assessment in China, Main Report and Technical Annex (Vol. 1–9) of Project CPR/87/029/B/01/R, United Nations Development Program, State Science and Technology Commission of the People's Republic of China, Food and Agriculture Organization of the United Nations, State Land Administration of the People's Republic of China, Beijing.

Wiberg, D., and Strzepek, K., 2000, CHARM: A Hydrologic Model for Land Use and Climate Change Studies in China, Interim Report IR–00-72, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Chapter 17

A Scalable Algorithm for Distributed Solution of Simulation-based Optimization in Groundwater Management

Manfred Grauer, Thomas Barth, Stefan Kaden, and Ingo Michels

Abstract

The solution of optimization problems is a common task in groundwater engineering, which is in turn an important discipline of environmental management. Exemplary optimization problems are the optimal design of a groundwater pumping installation, the determination of an optimal control strategy for it, or the optimal design of a discrete/digital controller used in level-controlled pumps. To evaluate the effects of human interference with a groundwater system, expensive simulations must be performed during the course of an optimization. This type of nonlinear optimization and control problem in groundwater engineering can be generally characterized as non-convex and non-smooth. Additionally, the involved simulation introduces a certain amount of "noise" to the solution process. These characteristics lead to very time-consuming solution processes, because of long-running computations for simulation and the difficulties in finding a global optimum. The basic concepts of a scalable algorithm for the solution of these problems are introduced.

The application of the distributed solution concept to optimization problems from groundwater engineering is shown using a parallel PVM-based implementation of the optimization algorithm and its integration with the groundwater simulation system FEFLOW®.

Keywords: Distributed optimization, simulation-based optimization, groundwater management, speedup, efficiency, scalability.

17.1 Introduction

Many complex optimization problems related to groundwater engineering cannot be formulated analytically. For constrained optimization and control problems, it is typical that the objective function and/or the constraints are highly nonlinear, and sometimes even have mixed-integer decision variables. Examples of typical problems are facility optimization problems in the water industry (Dandy *et al*., 1996; Jonoski *et al*., 1997). Due to the widely used simulations based on the Finite Element Method (FEM), the assumptions on convexity and smoothness of objective and constraint functions are not valid any more. Therefore, optimization algorithms with local convergence properties are not applicable. Furthermore, a global optimum of the problem must be found. These characteristics lead to an excessive computation time for the solution of a problem of this kind, making nonsequential solution approaches inevitable. Nonsequential algorithms can be used to reduce computation time by performing time-consuming computations in parallel, and they are also necessary to apply hybrid approaches to determine a global optimum (Boden and Grauer, 1995).

The computation of solutions to this class of problems typically requires one to perform numerically complex FEM simulations, which quite often have to be repeated many times during the course of a mathematical optimization. For example, simulation of a complex, three-dimensional groundwater model using FEFLOW®[1] (Diersch, 1998) takes about one hour of CPU time on a Pentium-III-based PC. In practice, optimization and FEM simulation are usually separate software systems, and thus lack the capability of mutually "calling" each other as subroutines. Thus, it is necessary to couple them as two "black boxes," because source-code-level integration is, in general, not possible.

Typically, the optimization module requests an FEM simulation for every evaluation of the objective function and/or the constraints of the optimization problem. The computation time of a single FEM simulation depends on the complexity of the simulated model and may range from a few seconds up to several hours.

[1]FEFLOW is a trademark of WASY Ltd., Berlin, Germany.

Many hundreds or thousands of these FEM simulations lead to very long computation times, a sequential strategy that sequential hardware architecture cannot cope with. The distribution of the entire optimization process on parallel computing hardware is inevitable. Instead of using "traditional" parallel hardware (MPP, vector computer), networks of workstations (NOW) can be used as a "virtual" parallel computer for distributed applications. This distribution can either be realized via parallel FEM simulation and/or parallel optimization algorithms. In both cases, data-parallel as well as task-parallel approaches for FEM simulation and/or optimization are appropriate. We focus on task-parallelism in FEM simulation, since domain decomposition methods executable on networks of workstations are not common in today's FEM simulation systems. Especially in the case of optimization algorithms, an approach integrating data-parallelism and task-parallelism is preferred. A step towards this integration is presented in this chapter by applying FEM simulators (task-parallel) for the computations requested by an optimization algorithm (data-parallel).

This optimization algorithm is a direct search method using only the information of the value of the objective function. The integration of task- and data-parallelism is realized by using a set of solutions (data-parallel) which are evaluated in parallel by multiple instances of a FEM simulator (task-parallel). To use the computational resources of a network of workstations, the optimization algorithm must be designed in such a way that almost all of the evaluations can be computed independently on the available workstations. A minimum of sequential components in the algorithm is desired to improve scalability and efficiency. These properties enforce the usability of a distributed algorithm on a network of workstations.

In this chapter, typical optimization problems from groundwater engineering are discussed in Section 17.2. A software architecture for supporting groundwater engineering based on the distributed computation of the simulation-based optimization problems is presented in Section 17.3. In Section 17.4, a distributed optimization algorithm for this problem class is briefly introduced. The computation of industrial problems is presented in Section 17.5 to show the feasibility of the approach for practical engineering problems. In Section 17.6, conclusions are drawn and areas for future research are discussed.

17.2 Classes of Optimization Problems in Groundwater Engineering

Three classes of mathematical nonlinear constrained optimization and control engineering problems will be presented. These classes represent typical optimization problems to be solved in groundwater engineering. Nevertheless, industrial applications of these different types of optimization problems can be found in various engineering domains: in the automotive industry (Boden, 1996; Weinert, 1994), in aircraft design, for metal forming processes (Grauer and Barth, 2001) and in chemical engineering (Grauer *et al.*, 1978).

In this section, $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ denotes the objective function of an n-dimensional optimization problem. The m equality constraints are given by $g_j(x) = 0, \forall j = 1, \ldots, m$ and k inequality constraints by $h_p(x) \leq 0, \forall p = 1, \ldots, k$.

17.2.1 The problem of optimal design (POD)

The set of feasible solutions $\mathcal{U}_{\text{POD}}$ to problems of the class of design or facility optimization is given by:

$$\mathcal{U}_{\text{POD}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{g}(\boldsymbol{x}) = 0, \boldsymbol{h}(\boldsymbol{x}) \leq 0\}.$$

The optimization problem can be stated as:

$$\{\min f(x) \mid x \in \mathcal{U}_{\text{POD}}\}\ . \tag{17.1}$$

This is the class of nonlinear constrained static optimization problems.

17.2.2 The problem of optimal control (POC)

Next, the class of optimal control problems is defined. The vector of time-dependent input variables of the control system consists of $\boldsymbol{z}(t)$ of so called "disturbance" variables and $\boldsymbol{u}(t)$ of q decision variables. The output vector is denoted by $\boldsymbol{y}(t)$ and the state vector is given by $\boldsymbol{s}(t)$. The dynamic behavior is described by a system of partial differential equations with start conditions and the given time horizon $M = [t_0, t_e]$. This problem forms the set $\mathcal{U}_{\text{POC}}$ of feasible controls:

$$\begin{aligned}\mathcal{U}_{\text{POC}} = \{\boldsymbol{u}(t) \in \mathbb{R}^q \ \mid \\ &\boldsymbol{g}(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(t)) = 0, \\ &\boldsymbol{h}(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(t)) \leq 0, \\ &\dot{\boldsymbol{x}} = \varphi(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(t), t), \text{and} \\ &\boldsymbol{x}(t_0) = \boldsymbol{x}_0, t \in M\}.\end{aligned}$$

The optimal control problem is then:

$$\left\{ \min \int_{t_0}^{t_e} \phi(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(t), t) dt \ \mid \ \boldsymbol{u}(t) \in \mathcal{U}_{\text{POC}} \right\}. \tag{17.2}$$

This control problem can – under certain conditions – be solved using the same algorithms as for solving POD, if it is transformed to a discrete time problem (Veliov, 1997). The time-discretization of the decision variables transforms a problem from class POC to a higher dimensional ($q * l$) problem of the class POD.

17.2.3 The problem of hybrid control optimization (PHO)

In contrast to the optimal control problem POC, the decision variables of hybrid control optimization problems do not depend directly on time. The time dependence is introduced via a closed-feedback loop with the time-dependent output variables $\boldsymbol{y}(t)$. This feedback loop realizes the control of a discrete system (e.g., digital control unit) over a continuous system (e.g., groundwater). The decision variables can be defined as $\boldsymbol{u}(\boldsymbol{y}(t))$ and the set of feasible solutions is:

$$
\begin{aligned}
\mathcal{U}_{\text{PHO}} = \{ & \boldsymbol{u}(\boldsymbol{y}(t)) \in I\!R^q \;\; | \\
& \boldsymbol{g}(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(\boldsymbol{y}(t))) = 0, \\
& \boldsymbol{h}(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(\boldsymbol{y}(t))) \leq 0, \\
& \dot{\boldsymbol{x}} = \varphi(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(t), t), \text{and} \\
& \boldsymbol{x}(t_0) = \boldsymbol{x}_0, t \in M\} \;.
\end{aligned}
$$

The optimization problem in this case is:

$$
\left\{ \min \int_{t_0}^{t_e} \phi(\boldsymbol{z}(t), \boldsymbol{s}(t), \boldsymbol{y}(t), \boldsymbol{u}(\boldsymbol{y}(t)), t) dt \;\; | \;\; \boldsymbol{u} \in \mathcal{U}_{\text{PHO}} \right\} . \tag{17.3}
$$

With the decision variables $\boldsymbol{u}(\boldsymbol{y}(t))$ being not directly dependent on time, this problem class can also be solved with the same algorithms as problems POD and POC.

17.3 Software Architecture for Distributed Simulation-Based Optimization

The problem classes presented in Section 17.2 and their solution processes yield various requirements for an integrated software environment supporting the solution of simulation-based optimization problems in engineering. The computationally expensive numerical solution process and the necessity to integrate FEM simulation packages with optimization essentially affect the software design. The following tasks for software engineering can be identified:

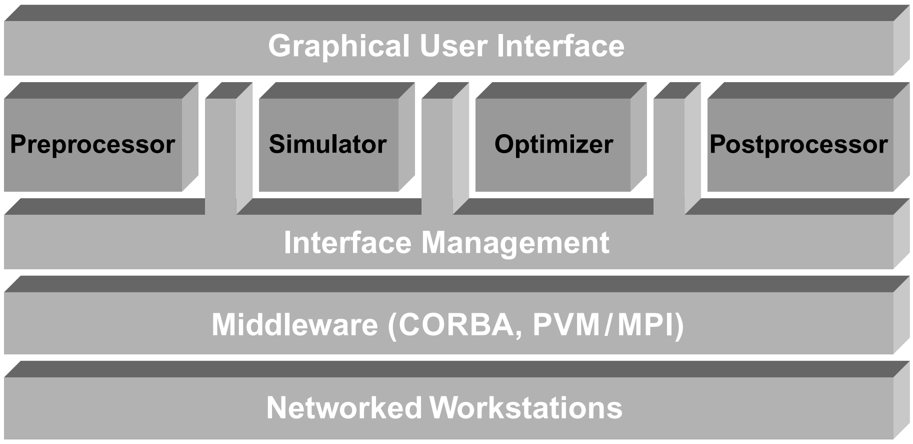


Figure 17.1. The software architecture for a distributed problem solving environment with coupled FEM simulation and optimization.

- **Wrapping of Legacy Systems** to preserve valuable domain-knowledge and enable seamless integration in newly designed object-oriented software environments,
- **Interface Management** to access and transfer data between simulation and optimization,
- **Synchronization** of control and data flow in simulation and optimization, and
- **Distributed Computation** to enable nonsequential solution approaches to reduce overall computation time.

A system architecture for coupling optimization and FEM simulation software systems is shown in *Figure 17.1*. The system we used to solve the introduced optimization problems is based on this architecture.

The two lower layers provide (platform-independent) functionality to start the individual components of the system and to provide the communication between them. These two layers implementing the middleware can be realized using an object-oriented approach like the Common Request Broker Architecture (CORBA) (Object Management Group, 1998) but alternatively, one could use the non-object-oriented Message Passing Interface standard (MPI) (Gropp *et al*., 1994), or the Parallel Virtual Machine (PVM) (Geist, 1994). The two lower layers also offer functionality for load distribution, e.g., via the WINNER resource management system for load distribution in networks of workstations (Arndt *et al*., 1998a,b; Arndt *et al*., 1999).

The layer above implements the interface management functions to provide the basis for application-level communication, i.e., exchange of values for decision variables and constraints between the components. This layer encapsulates the application-specific interface, whether it is file-based or a programming interface, and makes a common interface for data exchange available, e.g., by creation and/or transformation of files. Furthermore, synchronization between the components will be handled in this layer.

Components like the pre- and postprocessor have their own (graphical) user interfaces. The topmost layer has to provide a user interface for the convenient formulation of the optimization problem (e.g., by allowing nodes of a finite element model representing constraints or decision variables to be selected graphically, or by enabling the user-friendly specification of the objective function) and probably must also offer additional visualization techniques for the results of the optimization. As a whole, this layer should present the components of a coupled optimization and FEM simulation system consistently, and initiate and control the data flow between distributed components: from model generation in the preprocessing stage and FEM simulation/optimization to the visualization of the optimization results in the postprocessing stage.

17.4 The Scalable Optimization Algorithm

The design of a scalable distributed optimization algorithm for the solution of nonlinear constrained simulation-based optimization problems is substantially influenced by the experiences with other direct solution methods applied to such optimization problems (Grauer, 1987). In *Figure 17.2*, an overview of the general structure of the algorithm is given. A more detailed explanation of the algorithm's phases is provided by Barth *et al*. (2000a,b). In the following, the three phases are briefly described (n denotes the problem's dimension, p the number of available processors in the network):

1. **Initialization**
   The starting polytope is formed by $s > n + 1$ randomly generated solutions, where s is the number of points of the polytope. Constraints are evaluated on p workstations in parallel. Infeasible solutions are repaired using a parallel binary search directed toward the weighted center of gravity (Grauer, 1987) of feasible solutions.

2. **Exploration**
   The $e \leq s$ worst solutions of the s points of the polytope are reflected on and simultaneously moved toward the weighted center of gravity. Each of these reflection/contraction operations is performed l times in parallel. All

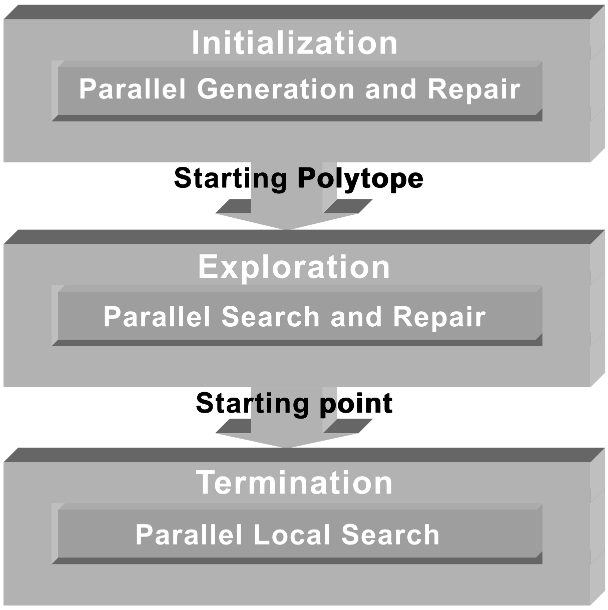


Figure 17.2. Schematic overview of the three phases of the distributed polytope algorithm for simulation-based optimization.

solutions are evaluated on p workstations in parallel. Infeasible solutions are repaired using parallel binary search. The s solutions for the polytope of the next iteration are selected from these solutions.

3. **Termination**
When the exploration is finished (e.g., after a given maximum number of iterations), a parallel local search starts from the best solution. It evaluates in parallel p random solutions in an environment around the best solution. The size of the search space is reduced if the local search fails to find a better solution. Infeasible solutions are rejected instead of repaired as in the previous phases. The local search stops after a given number of iterations or when the improvement is less than a given stopping criterion.

The exploration phase is the core part of the algorithm. The essential parameters for adapting the algorithm to problem size n and resources p determine the search strategy in this phase. Therefore, the scalability analysis in Barth *et al*. (2000b,c) is focused on these parameters. The scalability of the algorithm is demonstrated by means of actually measured applications presented in the following section.

By combining a more global search phase – exploration – with a local search, the algorithm follows the hybrid optimization approach to increase the probability of finding a global minimizer of the problem. In general, the distributed algorithm is able to evaluate more solutions in the search space in the same time compared to a sequential algorithm. This can also improve the quality of the solution by a better exploration of the search space.

17.5 Applications and Results

In the following sections, an overview is given of optimization problems related to groundwater engineering which are currently being investigated (Grauer *et al.*, 1999). Each of these problems can be associated with one of the problem classes introduced in Section 17.2. The distributed solution to these problems is based on the software architecture presented in Section 17.3. Results of the distributed optimization are provided for various problems in groundwater engineering. The first experiences show the importance of an adequate software architecture to provide an interface between optimization and simulation code.

17.5.1 Facility optimization in groundwater management

The following optimization problem from groundwater management is an example for the class POD defined by (17.1). It is a minimization problem with limits on the acceptable rise of groundwater level. The problem can be stated as follows. The process of building a sluice in a local port in a city causes an increasing infiltration of surface water and therefore a rise of the groundwater level. To protect the tree population in a nearby park, this rise of the groundwater level has been restricted to 0.1 meters in each of five observation points. The objective function of the minimization problem is the sum of the quantity of water (as a measure for the operational costs) that four pumps extract from the area in order to lower the groundwater level.

Formally, the optimization problem has:

- four decision variables (the individual quantities of extracted water per day of four pumps),
- five (implicit) constraints (upper bounds of groundwater level in five observation points), and
- four (explicit) constraints (technical restrictions of the pumps).

The problem was solved using the proposed software architecture with the FEM simulation system FEFLOW® (Diersch, 1998), a sequential Complex Box algorithm (Grauer, 1987), and the distributed polytope algorithm described in Barth (2000a,b,c) as the optimization algorithm. All computed solutions yield an improvement of approximately 25% compared to the reference solution of 1,600 cubic meters per day (m^3/day), which was manually determined by a parameter study performed by domain experts.

The distributed algorithm was performed on a varying number of hosts. In *Figure 17.3*, the relative speedup (Kumar *et al*., 1994) of the method is shown compared to theoretical results (for the complete analysis see Barth *et al*., 2000c) predicting the relative parallel speedup and efficiency of the algorithm. The corresponding efficiency is depicted in *Figure 17.4*. It can be seen that theoretical results are close to measured speedup and efficiency curves.

17.5.2 Optimal control in groundwater management

In this section, optimal groundwater management is described as an example for optimal control problems of class POC defined by (17.2) actually applied by the water industry. The problem is to find the optimal control strategy over a period of four years for one well with a maximum groundwater level in one observation point. In the observation point, the groundwater must be below a given level to protect several buildings. As a boundary condition, a groundwater recharge rate is given (see *Figure 17.5*). The control problem is transformed into a problem with six time intervals and a constant pumping rate for each of the intervals. Therefore, the resulting optimization problem has six decision variables.

The improvement of the distributed computed solution compared to the reference strategy is about 20% (see *Figure 17.6*). This reduction can be explained by the fact that the optimal strategy reduces the pumping rate when it is not necessary to extract water (in *Figure 17.6* approximately between day #1000 and day #1100). This allows the groundwater level to rise close to the upper limit but not beyond it.

The efficiency of the distributed algorithm is depicted in *Figure 17.7*. Due to the problems involved in the analysis of the algorithm's parallel speedup and efficiency concerning constraint handling (see Barth *et al*., 2000b,c), the actually measured efficiency is compared to different scenarios where the number of infeasible solutions v and the average number of repair operations is varied. Clearly, the measured efficiency is sufficiently close to the scenarios with two infeasible solutions ($v = 2$) per iteration and three repair steps ($q = 3$) for each of these infeasible solutions.

An explanation for superlinear efficiency for $p = 2$ and $p = 6$ is the small number of runs these diagrams are based on. In these cases, the randomly generated starting polytope contains a better solution than the one obtained by the sequential algorithm. We expect to further improve the sequentially computed solution for

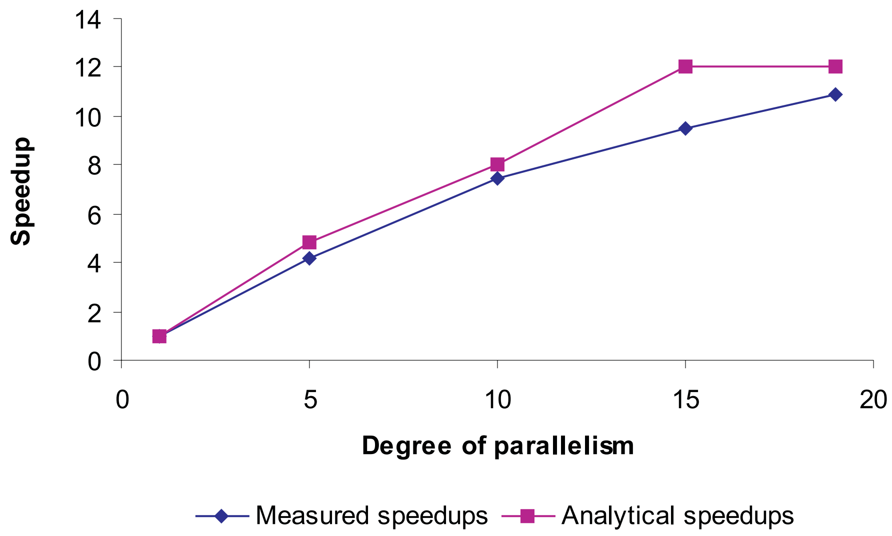


Figure 17.3. Speedup of the distributed polytope optimization algorithm with an increasing degree of parallelism for the solution of the facility optimization problem.

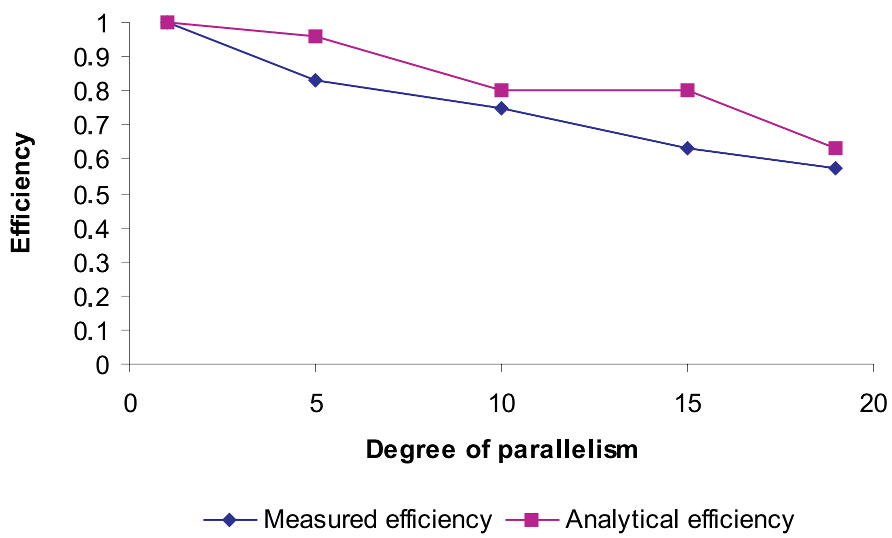


Figure 17.4. Efficiency of the distributed polytope optimization algorithm with an increasing degree of parallelism for the solution of the facility optimization problem.

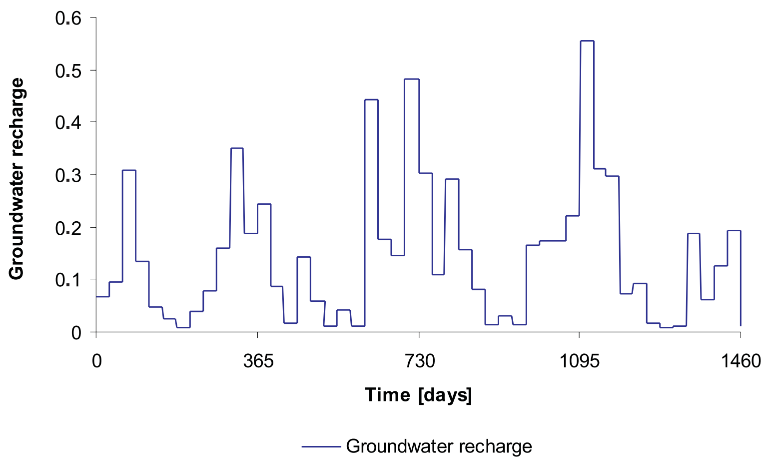


Figure 17.5. Groundwater recharge as disturbance $z(t)$ of the optimal control [Equation (17.2)] and the hybrid control problems [Equation (17.3)].

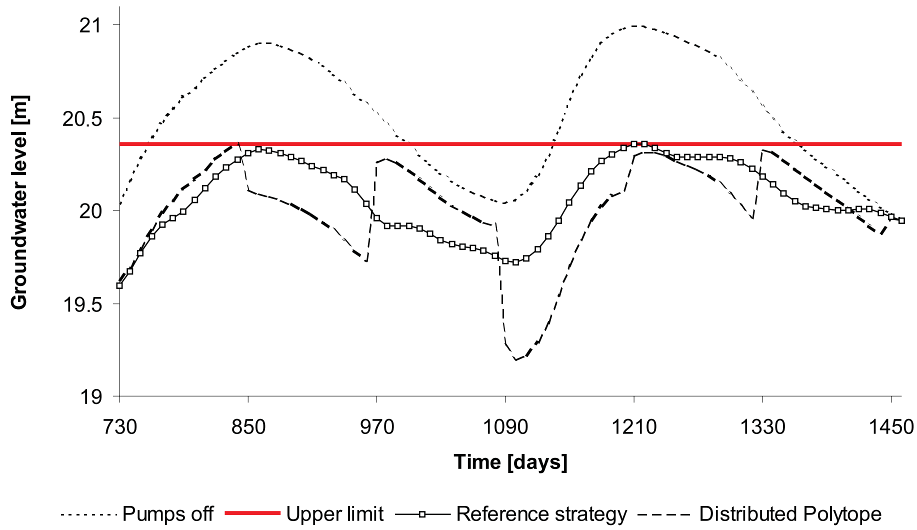


Figure 17.6. The trajectory of the groundwater level in one observation point as a result of the optimal strategy solving POC computed by the distributed polytope method. Compare results with the upper limit, the reference strategy, and the trajectory with all pumps switched off.

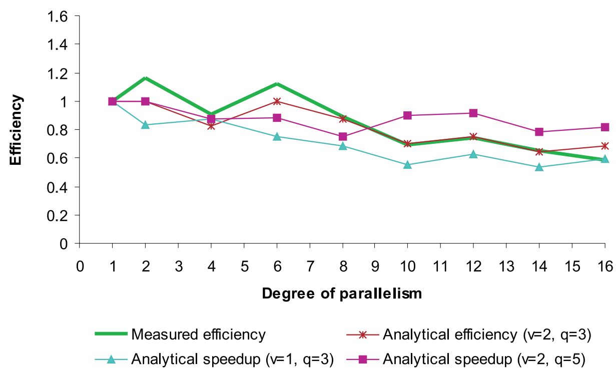


Figure 17.7. Efficiency of the distributed polytope optimization algorithm for the solution of the optimal control problem POC in groundwater management compared to different scenarios (varying parameter v and q) for the search space analysis.

each degree of parallelism when performing more optimization runs. In general, speedup and efficiency are not necessarily strictly increasing functions due to the properties of the distributed algorithm. In the case of $p = 20$ available hosts, 18 parallel simulations can be performed in the same time as 16, 17, or even 19. This means that the speedup remains constant in these cases and efficiency is decreasing. Adjustment of parameters of the algorithm is responsible for an efficient relation between the number of hosts and the number of simulations computed in parallel.

17.5.3 Hybrid control in groundwater management

The scenario which is the background for the problem of hybrid control problems of class PHO defined by (17.3) is as follows: A well extracts a quantity of water to keep the groundwater level in given observation points below a certain maximum, e.g., to protect basements from water damage. This well is switched on and off by another well: this control well switches between different water levels (quantities) depending on the measured groundwater level. This scenario is depicted in *Figure 17.8*. The groundwater recharge rate is again given as a boundary condition (see also *Figure 17.5*).

The optimization problem has five decision variables (groundwater levels where the pump switches between pumping rates). In *Figure 17.9*, the optimal trajectory of the groundwater level in the observation point (using the optimal switching levels) is shown. The activities of the pump using the optimized levels for switching is shown in *Figure 17.10*. The optimized levels improved the reference solution by almost 13%.

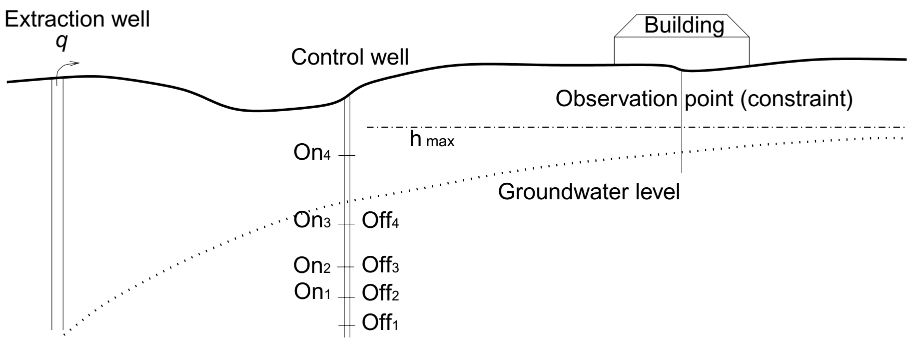


Figure 17.8. Scenario for a hybrid control optimization problem PHO in groundwater management. The control well switches the quantity q of the extraction well depending on the switching levels On_i and Off_i to maintain an upper limit $h_{\max}$ on the groundwater level.

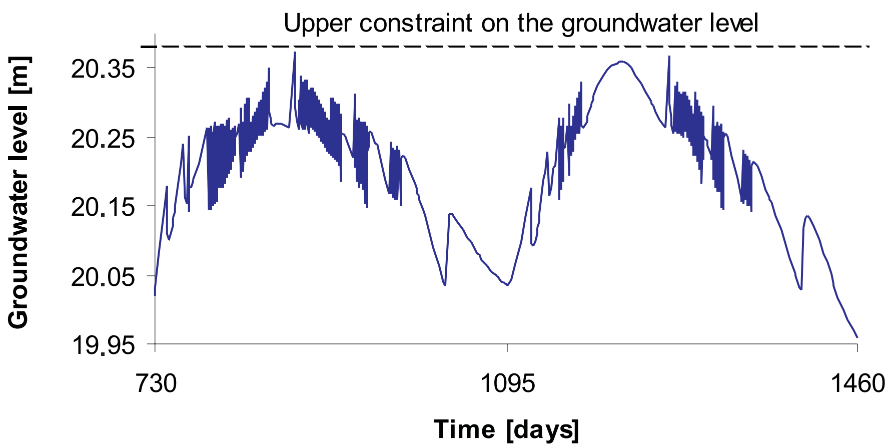


Figure 17.9. Groundwater level in the observation point using the optimal switching levels as a result of the solution of the hybrid control problem PHO.

17.6 Conclusions and Future Work

In this chapter, three classes of optimization problems were presented from groundwater engineering and the architecture of software environments suitable to solve these problems on networks of workstations. An approach for the distributed computation of numerically complex, long running optimization problems in engineering applications was described. The feasibility of both the software architecture and the distributed solution approach was demonstrated. Furthermore, speedup

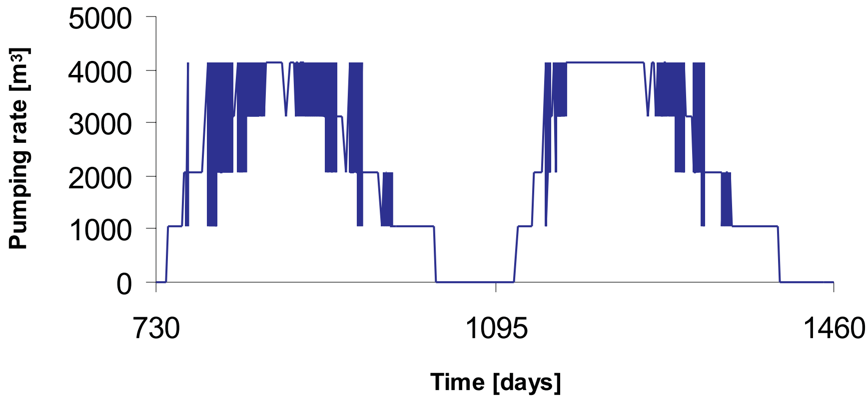


Figure 17.10. Pumping rates $u(t)$ (in m^3/day) of the extraction well using the optimal switching levels as a result of the solution of the hybrid control problem PHO.

and efficiency of the proposed method was validated by applying it to instances of the presented problem classes.

There are several issues for future work. We plan to solve problems from other engineering domains like aircraft and automotive design and production. Thus, the interface management component of the architecture is the subject of future work. Furthermore, as indicated by the performance results, the efficiency of the distributed optimization algorithm should be improved. We envision the adaptation of its search strategy according to information gathered during the optimization, e.g., by adjusting parameters of the algorithm. A convergence analysis and considerations on the complexity of scalable, distributed optimization algorithms is also in the focus of future research. Finally, as an alternative for distributed computing based on PVM message passing, the use of CORBA (equipped with adaptive load distribution) as a middleware is currently being developed (Barth *et al*., 1999).

Acknowledgments

This research is partially funded by the German Federal Ministry of Science and Technology under contract FUEGO #0033701E8. Furthermore, the authors would like to thank H. Diersch and R. Gründler from WASY Ltd., Berlin for their cooperation and support.

References

Arndt, O., Freisleben, B., Kielmann, T., and Thilo, F., 1998a, Dynamic load distribution with the WINNER system, in *Proceedings of the Workshop Anwendungsbezogene Lastverteilung* (ALV'98), SFB 342/01/98, University Press, Munich, Germany, pp. 77–88.

Arndt, O., Freisleben, B., Kielmann, T., and Thilo, F., 1998b, Scheduling parallel applications in networks of mixed uniprocessor/multiprocessor workstations, in *Proceedings of Parallel and Distributed Computing Systems* (PDCS98), International Society for Computers and Their Application, Chicago, IL, USA, pp. 190–197.

Arndt, O., Freisleben, B., Kielmann, T., and Thilo, F., 1999, Batch queueing in the WINNER resource management system, in *Proceedings of the 1999 International Conference on Parallel and Distributed Processing Techniques and Applications* (PDPTA'99), CSREA Press, Las Vegas, NV, USA, pp. 2523–2529.

Barth, T., Flender, G., Freisleben, B., and Thilo, F., 1999, Load distribution in a CORBA environment, in *Proceedings of the International Symposium on Distributed Object and Application 99*, IEEE Press, Edinburgh, Scotland, UK, pp. 158–166.

Barth, T., Freisleben, B., Grauer, M., and Thilo, F., 2000a, Distributed solution of simulation-based optimization problems on networks of workstations, accepted for publication at the 4th Workshop on Runtime Systems for Parallel Programming (RTSPP), Cancun, Mexico.

Barth, T., Freisleben, B., Grauer, M., and Thilo, F., 2000b, A scalable algorithm for the solution of simulation-based optimization problems, in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications* (PDPTA'2000), CSREA Press, Las Vegas, NV, USA.

Barth, T., Freisleben, B., Grauer, M., and Thilo, F., 2000c, Distributed solution of optimization problems and the design of optimal hybrid control systems on networks of workstations, accepted for publication at the IEEE International Conference on Cluster Computing (CLUSTER 2000), 28 November–2 December, Chemnitz, Germany.

Boden, H., 1996, *Multidisciplinary Optimization and Cluster Computing*, Springer/Physica Verlag, Heidelberg, Germany (in German).

Boden, H., and Grauer, M., 1995, OpTiX-II: A software environment for the parallel solution of nonlinear optimization problems, *Annals of Operations Research*, **59**:129–140.

Dandy, G.C., Simpson, A.R., and Murphy, L.J., 1996, An improved genetic algorithm for pipe network optimization, *Water Resources Research*, **32**(2):449–458.

Diersch, H.-J., 1998, *Interactive, Graphics-Based Finite-Element Simulation System FEFLOW for Modeling Groundwater Flow Contaminant Mass and Heat Transport, Users Manual*, WASY GmbH, Berlin, Germany.

Geist, A., 1994, *PVM: Parallel Virtual Machine – A Users Guide and Tutorial for Network Parallel Computing*, MIT Press, Cambridge, MA, USA.

Grauer, M., 1987, Optimierung verfahrenstechnischer Systeme, in S. Weiß *et al*., eds, *Verfahrenstechnische Berechnungsmethoden*, Part 6, Section 4, Verlag Chemie, Weinheim, Germany, pp. 127–129 (in German).

Grauer, M., and Barth, T., 2001, Cluster computing for treating MDO-problems by OpTiX, in F. Mistree and A. Belegundu, eds, *Proceedings of the Conference on Optimization in Industry II*, ASME, New York, NY, USA.

Grauer, M., Barth, T., Kaden, S., and Michels, I., 1999, Decision support and distributed computing in groundwater management, in D. Savic and G. Walters, eds, *Water Industry Systems: Modelling and Optimization Applications*, Research Studies Press, Baldock, Hartfordshire, UK.

Grauer, M., Gruhn, G., and Richter, C., 1978, Design and application of the GRG-algorithm for process optimization, 11th CACE '78, Paris, France.

Gropp, W., Lusk, E., and Skjellum, A., 1994, *Using MPI: Portable Parallel Programming with the Message-Passing Interface*, MIT Press, Cambridge, MA, USA.

Jonoski, A., Zhou, Y., Nonner, J., and Meijer, S., 1997, Model-aided design and optimization of artificial recharge-pumping systems, *Hydrogeological Science Journal*, **42**(6):937–953.

Kumar, V., Grama, A., Gupta, A., and Karypis, G., 1994, *Introduction to Parallel Computing – Design and Analysis of Algorithms*, Benjamin/Cummings Publishing, Redwood City, CA, USA.

Object Management Group, 1998, *The Common Object Request Broker: Architecture and Specification*, Revision 2.2, Object Management Group (ftp://ftp.omg.org/pub/docs/formal/98-07-01.ps).

Veliov, V., 1997, On the time-discretization of control systems, *SIAM J. on Contr. & Opt.*, **35**(5):1470–1486.

Weinert, M., 1994, *Sequential and Parallel Strategies for the Optimum Design of Complex Shells of Revolution*, TIM-Report T05-05.94, University Press, Siegen, Germany (in German).

Chapter 18

Changes to the Polish Water Management System

Jolanta Pakulska

Abstract

Until recently, water management in Poland was based on the administrative division of the state. This situation did not provide good results in the protection of water resources and made rational water management difficult. To cope with this, the Ministry of Environment proposed a new system of water management in the new Water Law that was based on regional catchment areas. The law created a Regional Board of Water Management, which is responsible for implementing national strategies for water management. The new Water Law gives local governments more power, although they will still have to report on their water management activities to the Regional Board for Water Management. The law allows regionalization of water resources management and gives local governments more autonomy. It also includes new economic instruments that affect the behavior of polluters.

Keywords: Water management, aim, rules, catchment area, sustainable development, economic instruments for water management, water law, underground water, surface water.

18.1 Introduction

The preservation of Polish water resources began with the first Water Act, issued on the 19th of September, 1922. It was based on laws designed by the invaders: the Austrian one (from 1875) and the Prussian one (modified at the beginning of the 20th century). Among other things, the law concerned:

- the ownership and exploitation of water,
- the maintenance of water resources and their banks,
- flood protection, and
- the organization of water companies.

The law did well at regulating issues concerning water resources and effectively protected water from being polluted. After World War II, changes made it necessary to restructure water authorities, and frequent transformations occurred. For a long time, the Ministry of Environmental Protection, Natural Supplies, and Forestry was the principal State organ. In 1985, this organization was transformed into the Ministry of Environmental Protection and Water Economies. Following the most recent organizational changes of the Polish government (1998), responsibility for water resources is currently lodged with the Ministry of Environment. Its sphere of activity is:

- preservation of the environment, including protection of water and atmospheric air;
- water management;
- maintaining rivers and streams; and
- flood protection.

18.2 Present System of Water Management

18.2.1 The aims of water management

The State affords Polish water resources, as a part of the overall environment, a special sort of protection. As a result, the authorities undertake actions aimed at preserving the environment. The preamble to the environmental law dated the 31st of January 1980 emphasizes that in order to ensure preservation and rational shaping of the environment (including water resources) it is vital to:

- shape the environment according to the needs of the national economy;

- harmonize economic development with the needs of environment protection;
- define the law to ensure that the environment is used in compliance with social needs;
- make environmental protection tasks an obligation of State, collective, social, and professional institutions; and also
- make the protection of the environment compulsory.

The management of water resources encompasses preventing water pollution and ensuring rational water management. Pollution prevention means that the level of water quality should be maintained in accordance with certain regulations. It is necessary to prevent any harmful pollution from bringing about changes in the physical, chemical, and biological properties of water, changes that make water useless for the needs of people and the national economy and that cause harmful modifications in the environment (Water Law, unpublished proposal). The aim of water preservation in Poland is to:

- ensure rational use of water resources; and
- prevent or counteract changes to the waters, which make them useless to people, flora, fauna, or the national economy (Environment Law, 1980).

The purpose of water management in Poland is to satisfy water needs for people and the national economy. The basic rule is this: water usage must not cause water pollution. Underground water and areas feeding it are under special protection. In the Polish legislation there are three ways of using water, each of which is connected with various laws and duties. The three ways of using water are:

1. Public use, including satisfying personal needs and household needs without using special equipment. Public use covers entertainment, tourism, water sports, and angling.
2. Ordinary water use, which entitles the land owner to satisfy his/her personal needs and the needs of the agricultural household. The water is his/her own possession, as is underground water allocated on his/her land, and the owner needs no water law permission.
3. Special water use, which requires a water law permission, which is issued indefinitely and explains in detail the conditions under which the water resources can be used.

The general aim of water management is to satisfy justified needs connected with water usage in defined quantity and of defined quality, in a defined place and time (Łustacz, 1991). The amount of water in Poland is not large in comparison to that in countries of similar climatic circumstances. Renewable surface waters account for 55 cubic kilometers (km^3) (64 km^3 together with inflows from foreign lands). From this discretionary supply, 49.4 km^3 are available (of 56.2 km^3 together with inflows from foreign lands). Water resources in Poland supply 1,600 cubic meters (m^3) per inhabitant per year (the European average is a little over 4,000 m^3, and world over is 12,000 m^3). The magnitude of water supplies puts Poland in just a little better situation than Egypt. Poland is the only country in Europe known as a "water stress" state. Each inhabitant of a water stress state gets only 1,000–1,700 m^3 water per year.

Underground water supplies are also limited. Exploitable underground resources account for 13.7 billion m^3 per year, or 1.1 m^3 per inhabitant per day. Underground water meets about 16% of Poland's water needs.

Water supplies in Poland are bounded by many parameters. For the purposes of implementation, classifications are compared, and concentrations of pollutants with admissible magnitudes are set forth for three classes of quality according to two criteria: biological and chemical. Three classes of water quality are distinguished, according to the aims for each one. An aggregated coefficient of pollution is the coefficient of biochemical application of oxygen, which addresses water containing decomposing organic matter.

Over the last 10 years, water of the first class practically disappeared. Taking into account the quantity of water instead of the length of the river, the most polluted rivers are those richest in water, and their central and upper sections. The quality of Polish rivers got considerably worse. When we consider the quality of water in lakes, a similarly bad situation is seen. Large or average lakes contain water mostly of the third or second classes. Most lakes contain water of the second class, but those with the greatest volume have water of the third class – almost useless to ordinary use. The quality of underground water is higher than that of surface water, because it is isolated from external influences and is subject to special protection. Water that because of its high quality can be a source of supply for the population is over 99% of a whole public supplies. In this, more than 20% does not demand cleaning, and more than 65% is polluted only slightly and it is easy to make it useful.

The details of water quality are regulated by the order about water classification and standards that the sewage directed to water or soil must fulfill (see *Table 18.1*). Discharging sewage into waters must conform to the following requirements:

- no changes in water management or temperature,
- formation of no sediment or foam, and
- no changes in natural foam and turbidity or in natural biocoenosis typical of waters (Rule, 1991).

Table 18.1. Standards of selected sewage directed to water or soil.

Parameter	Standard
Temperature	35°C
Reaction	6.5–9.0 pH
Suspensions	50.0 mg/l
Biochemical application of oxygen (5 days)	30.0 mg O_2/l
Chemical application of oxygen	150.0 mg O_2/l
Organic carbon	40.0 mg C/l
Nitrogen	30.0 mg N/l
Phosphorus	5.0 mg P/l
Chlorides	1000.0 Cl/l
Sulfates	500.0 mg SO_4/l
Dissolved substances	2000.0 mg/l
Zinc	2.0 mg Zn/l
Copper	0.5 mg Cu/l
Nickel	2.0 mg Ni/l
Lead	0.5 mg Pb/l
Mercury	0.02 Hg/l
Silver	0.2 mg Ag/l
Fluoride	15.0 mg F/l
Sulfides	0.2 mg S/l

Source: Pakulska (1996).

The aims of water management in Poland are to:

- improve the situation in regions with water shortage;
- introduce practical, economical instruments that would improve the use of water and the environment (introducing rules of sustainable development); and
- establish the sensible use of water supplies and resources.

The deregulation of water resources considerably influences Polish water management. The issue is very significant to the priorities of Polish water management because, under deregulation, the Board's main purpose is to provide enough water for municipal and industrial use. This task is closely linked to the following problems:

- The irregular allocation of water causes acute lack of water in many regions of the country. Accumulation of water in storage reservoirs will not solve this problem (retention in Poland equals 5% of average outflow while the world's average is 11%). Furthermore, this low retention was further undermined by mismanaged drainage operations, which were based on one-way reclamation that devasted soil and forests.

- Industry uses water-intensive technologies, because there is too little economic motivation to invest in technology changes.
- Highly water-inefficient industries are frequently located in regions where water is scarce, which increases the problems of water shortage. Moreover, industry uses high-quality underground water for its purposes, while municipalities are forced to use surface water of much lower quality.
- Water quality is endangered by industry, agriculture, and forestry through the burning of sulfurized coals, the danger of salinity of mine waters, the drainage of mines leading up to the depletion of underground waters, lack of defined regulations concerning waste storage, irrational use of fertilizers and chemical means of plant preservation, the devastation of forests and deforestation, ill-considered drainage, inadequate consideration for the water–protecting function of forests, and deforestation of peatbogs and fens (Kleczkowski, 1984).

Therefore, management of water resources needs to be reorganized from the structure based on administrative borders to the structure based on catchments.

18.2.2 The rules of water management

From the beginning of the 1990s, management rules were formulated consistently and coordinated with the ecological policy of the state. In 1991, the Ministry of the Environment, Natural Resources and Forestry designed a new ecological policy for the country, which also covered water resources. This policy was based on the concept of sustainable development and was updated in 1999. This policy requires that resources and values of the environment should be kept in a condition that allows their exploitation not only by the present generations but the future ones. This principle requires equal treatment of social, economic, and ecological goals and is composed of the following elements (principles):

1. Equal access to the environment.
2. Regionalization.
3. Socialization.
4. Polluter pays.
5. Prevention.
6. Far-sightedness.
7. Using only the best available solutions.
8. Complementarity.
9. Ecological and economic effectiveness.

In Poland's new ecological policy, market instruments play an important role. Their use is based on the following principles:

1. Economization, i.e., achievement of ecological goals at minimal cost.
2. Conforming to the lawfully and administratively agreed-upon level of interference in the environment.
3. Regionalization, i.e., adjusting protection requirements to regional conditions and enabling the local authorities to establish the local instruments of ecological policy.
4. Financial responsibility of the polluter, who has caused environmental damage, for the costs of reclamation.

The principle of economization means, in practice, that the players whose costs are the lowest ought to conform to the most stringent requirements. Fulfilling principles (2) and (3) raises a lot of problems. Standards for pollution emission are not often suitable because there is no threat of major consequences. Also, the sphere of regionalization of water management is still limited. Some steps have been taken toward realization of principle (4), e.g., the fees for using the environment and the penalties for violating standards are now at a more rational level.

18.2.3 Premises of the planned system change

The aims and rules of water economy in Poland are already in line with modern management systems established in the countries of the European Union. They are also consistent with the rules of sustainable development.

A rational water management structure aims at meeting the water demands of all users, and to balance water resources and supply with the demands. This is why regional differentiation of water management instruments is so vital. One must take irregular allocation of surface and underground water across the country into account.

The system of water resources management should take into consideration the depletion of resources and abandon treating them as a free, unlimited resource. It ought to encourage all economic units and individual users to use water resources in an economical way. The principles of using resources should be constructed in such a way that the demand would be met by supply. Economic changes combined with the changes in the system of water management should contribute to the increase in effectiveness of policy in proecological activities.

The basic element of the system of water management is the Water Law (unpublished proposal). The current management system, based on the administrative division of the state, cannot ensure effective water resources management. Work toward reforming that system has been carried out at the Ministry of the Environment since the beginning of 1990s, and is modeled after the French scheme based on catchment areas. Basically, all decision makers agree that such changes are necessary. However, so far we lack a process of decisionmaking for actual implementation of the Water Law.

18.3 The New Water Management System

18.3.1 Organization of water management

The system proposed by the Ministry of the Environment is based on the following assumptions:

1. Turning from management based on administrative units to management by hydrology, i.e., in water–economical regions.
2. Separate administrative and control activities from economic activity.
3. The assumption that water needs protection from pollution should be an integral element of water resources management (Kozłowski, 1991).

Within the limits of the reform of the management system into catchment areas, seven Regional Boards for Water Management have been set up (see *Figure 18.1*). They have legal status and the Minister assigns their director. Each board will be responsible for fulfilling policies in the catchment assigned to it. Each catchment will have its own Water Management Council, consisting of representatives of administration, people elected to municipal governments, and water users.

Regional Boards for Water Management design conditions for using catchment water, which, after being discussed with and agreed upon with their Water Management Council are approved by the Minister of the Environment. Moreover, the Boards collect fees for special use of water, as well as water-related materials in the catchment for water economy purposes, 15% of which will be donated to the central fund to finance tasks of water economy having strategic meaning from the point of view of the state (Koza, 1994). Duties of the Catchment Council will include establishing programs of water economy in the catchment, assigning the raised funds, and defining fees for using water and related materials in the catchment.

Thanks to this reorganization, special conditions for regionalization of fees will be created, and at the same time conditions for regional systems of financing water economy. The government will, as in the past, determine basic rates, while the Councils will be able to differentiate them freely. The new Water Law preserves the organization of water usage into the traditional categories: public, ordinary, and special. It has also kept the idea that only the one category requires water-law

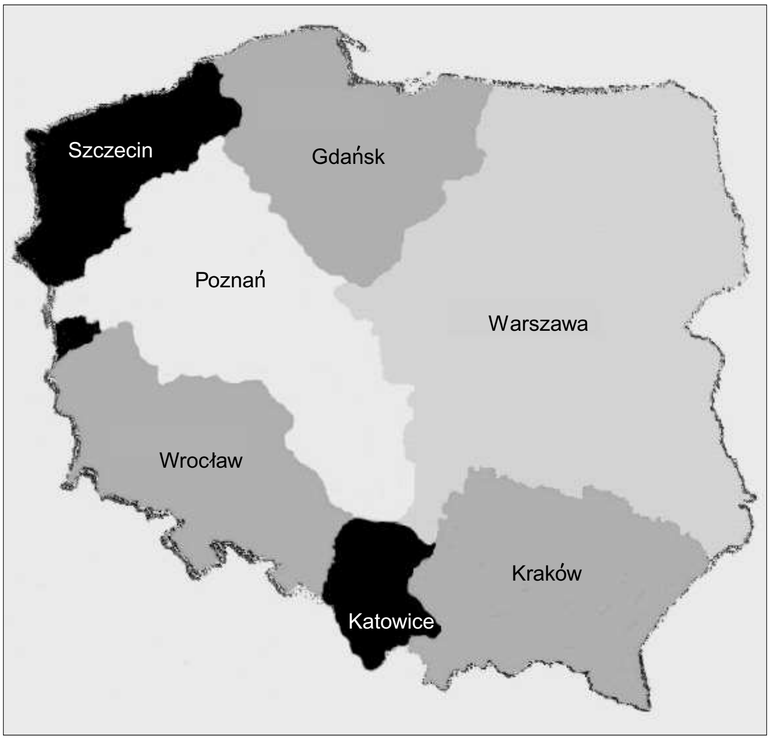


Figure 18.1. Regional catchment areas. Source: Pakulska (1998a).

permission. The basic element of the system is still to be the water-law permission for special use of water, which is to be a sort of license for using water.

Regional Boards for Water Management were appointed in February 1991. At this time, the Boards were to design the conditions of using catchments and to create an information system of water management for catchments. The appointment of Regional Boards was necessary to carry out preparatory work. Today, the Regional Boards are prepared to take over their duties but there is a lack of legal basis to implement it.

18.3.2 Economic instruments for water management

Economic instruments make up important aspects of water resources administration. For the market economy, these instruments create a very important impetus than can stimulate economic agents to actions that protect water resources. Economic impetus enables indirect financial influence on agents that gives them simultaneous choices (for instance, paying fees for draining of sewage or building a sewage-treatment plant and bearing the costs of its use).

In Poland, many economic tools of water management are used. Their deployment induces economic agents to undertake domestic activities that protect the environment. The most important are ecological payments, and among these are favored:

- payments for polluting the environment, which depend on quantities and qualities emitted, i.e., discharging sewage into waters; and
- economic payments for profiting from the environment, e.g,. payments for use of water or for taking materials (e.g., sand, stones, gravel) from waters.

There are also financial impulses for application of the law. Most important among them are penalties for disturbing the environment (e.g., for not conforming to ecological rules). Ecological deposits are also used to stimulate observation of the rules.

Economics also enable policy makers to create a system that discourages polluters from violating set standards and stimulates them to undertake desired protective actions (Pakulska, 1998b), but the fees must be established at the correct level (see *Figure 18.2*).

Environmental Protection Fees

Environmental protection fees are like a price paid for the use of the environment. Correct fees for water consumption are very important to the rational usage of water. Nowadays in Poland all economic entities are obligated to pay in advance for consumption water when they use more than 10,000 m^3 a year. The rate depends on the kind of water used (surface or underground water), as well as on the region of the country and the purpose (see *Table 18.2* and *Table 18.3*). Regional differentiation of fees depends on the size of water resources. The higher rates for regions of water shortage were based on the average costs of increasing available water resources (building reservoirs, water transfers, etc.). In areas with enough water, the rate is fixed at a lower level and is based on the costs of servicing river-regulating buildings (Symonowicz, 1988). There is no differentiation in regional rates for water used for municipalities and agriculture. This water is far cheaper than that used for other purposes.

The differentiation of water prices aims at motivating users to spare it. That is possible only when the price of water is higher than the costs of saving it (e.g., by switching to a water-saving technology). In order to fulfill this condition, the rate should be based on economical analysis. On the basis of such analysis, the prices of water can be set for various industrial branches. Moreover, the size of the region's water resources should be considered in order to vary the rate. Formation

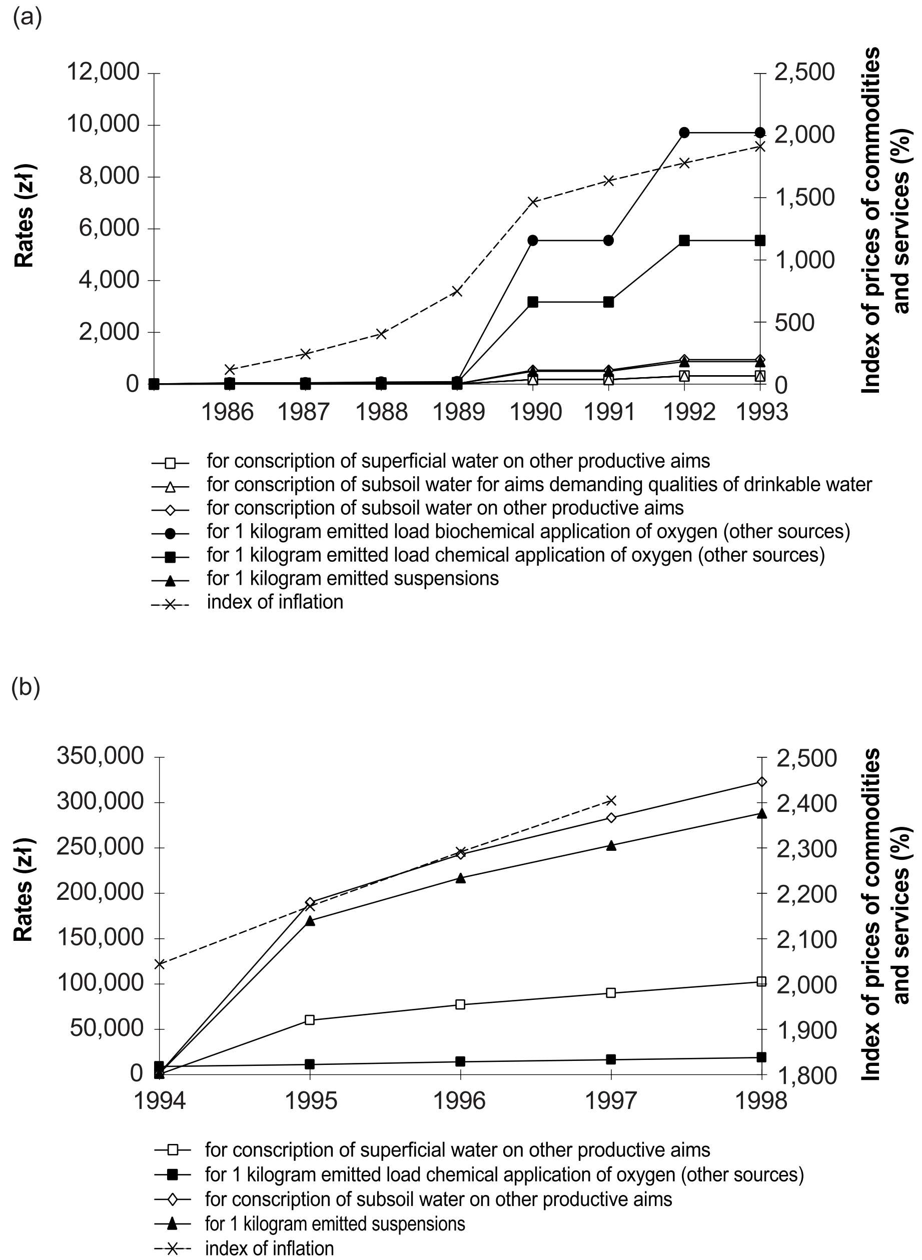


Figure 18.2. Height of select rates of fees and index of inflation in Poland in years (a) 1975–1993 and (b) 1994–1998. Source: Pakulska (1998b).

Table 18.2. Multiplier payments for use of surface water.

Province	Use	Multiplier
Group I (south of Poland)	Production of energy	0.5
	Other	2.5
Katowice	Production of energy	1
	Other	5
Other	Production of energy	0.2
	Other	1

Source: Based on Rule, 1991.

Table 18.3. Multiplier payments for use of underground water.

Use	Area	Multiplier
Productive purposes, in which water enters in composition or immediate contact with food or pharmacological products	Group I	0.8
	Other	0.3
Other purposes of production	Group I	2.5
	Other	1.0

Source: Based on Rule, 1991.

of Regional Boards will enable a better adjustment of the above-mentioned rates to the local situation.

At present, the instrument used most frequently in Poland is the fee for emission of pollution into natural water systems. This fee was put into practice in the 1970s. These charges are based on the water laws and other rules and regulations of Environment's Department of Provincial Offices (and on individual standards for emission), and they are included in the costs of the operator's activity. The base for setting the fees is the quantity of emission and its toxicity: the highest amounts are applied to metallurgy, machines, and electrical engineering, and chemical and light industries, the lowest rates are for municipalities and for hospitals. The rates for these do not reflect damage done by discharging sewage (Pakulska, 1998b).

The proposed new Water Law contains a regulation forbidding drainage of dirty sewage, which will in turn result in limiting the quantity of its emission. Sewage has to be cleaned to the prescribed level before it is discharged into water or into soil, and the discharge should meet the standards defined by the law. One decisive factor that would make this requirement effective is its sufficiently high standards defined on executive file.

Apart from gauging the rates of environmental protection fees correctly, it is essential to have efficient and effective services for inspection and control-measuring. In Poland, about 20% of economic entities listed in the statistics of the Central Bureau for Statistics do not employ any staff in the water-supply-and-sewerage

Table 18.4. Multiplier of payments for discharges of polluted sewage.

Purpose	Biochemical application of oxygen	Chemical application of oxygen
Chemical industry, petrol and power industry, metallurgical, electromechanical, light	2.00	2.5
Wood and paper industry	0.85	0.9
Food industry	0.5	0.6
Water-supply-and-sewerage devices	0.2	0.2
Other sources	1.00	1.0

Source: Based on Rule, 1991.

economy, and many of them employ an inadequate number of such staff. Supervision performed by special environmental services is also insufficient.

The main aim of environmental protection fees is to stimulate economical use of shrinking resources. Therefore, the environmental protection fees are accumulated into special funds, and pro-environmental investments are also funded and subsidized. One possibility for fixing the fee rate at the proper economic level is based on an accounting of costs that reflects all damages to the environment, and setting rates at a level that covers costs for environmental protection and reclamation. Raising the rates to such levels is not implementable at present, because only few economic entities would be able to cope with them.

Ecological Penalties

The ecological penalty is a financial stimulus to observe regulations and comply with the requirements of law. A provincial sanitary inspector of emissions into natural water systems imposes the penalties, which, however, do not correspond to the proposed requirements yet. The penalties are not based on the evaluation of environmental damages, but are based on the emissions above an approved amount.

Ecological penalties fulfill a disciplinary function. The mode of administering penalties is similar to that of administering environmental protection fees and their size depends on an evaluation of the damage caused by discharging polluted water (see *Table 18.4*). Such penalties decrease the profit of the offending entity. Like protection fees, they are transferred to ecological funds.

In the new Water Law, the term “ecological penalties” is replaced by “progressive fees.” They will be fixed dependently on the water used without permission or over the allowed quantity and for emission of pollution without permission or over a fixed level. The rates of these progressive fees can be, at a maximum, 10 times bigger than the ecological ones. They have the character of penalties and are calculated by the same entity that establishes environmental protection fees (that is, by Provincial Office of Environment).

Unlike penalties, which existed up until now, it is impossible to appeal the decision that fixed the fine and to pay by installments, both of which are characteristic for protection fees. Most polluters appeal decisions imposing a penalty, and the appealing procedure can last a very long time. Under this scheme, even when the appeal is not successful – and very often it is – the penalty loses its power of influence. Replacing these penalties by progressive fees will also result in the polluter paying statutory interest for the delay if payment is not made on time. These fees will not be charged, like present penalties, against the profit of the enterprise, but will rather be included in the expenses of the activity. Thus, they will be transferred into the price of the product. Competition on the market will prevent an unlimited rise in prices. Progressive fees should, therefore, better meet the expectations concerning stimulating the desired behavior of users.

Other Instruments

In addition to the changes in function of old economic instruments, policy makers are planning to introduce new instruments, for example:

- product environmental protection fees,
- ecological insurance, and
- licenses (emission permits).

So-called product environmental protection fees will be applied for products especially burdensome to the environment, which are used in huge quantity and in a diffuse way (for example, ecological overheads on fuel). Rates should satisfy two functions: they should act as a stimulus and as a redistributing force (accumulating money on ecological purposes). Increased process of products should encourage consumers to choose products less harmful to the environment, and simultaneously cheaper. On the other hand, they assure realization of the Polluter Pays Principle. It is being planned to make these charges cover nitric and phosphoric fertilizers and pesticides (which seriously endanger surface waters) as well as empty packages and washing powders, which contain a lot of phosphorus.

Ecological insurance will provide risk transfer opportunities, particularly for entities running activities having substantial environmental risks.

18.4 Conclusions

The administration of water management based on catchment areas seems to be the best solution for Poland, because it will ensure a regional perspective and self-government of local environments. Granting this regional perspective means transferring some entitlements from the State to the local administration. The State only plays a role of coordinator for transregional and centralized undertakings. Moving over to the system entailing the Regional Catchment Council will increase the entitlements for self-governments at the cost of the province's authorities, and some of the areas of authority will be taken over by the Regional Boards for Catchment Areas. The problems connected with administration of water resources are of a local nature, and can be taken note of and solved more effectively by self-governing organs, than by the State administration. Transfer of the entitlements to Regional Boards and to Councils enables the players to differentiate economic mechanisms and environmental politics (if it is possible to arrange regional fees, standards, etc.).

One can raise many objections about the proposed Water Law. But it is necessary to state that it is up to date and based on the most modern solutions applied by the West European countries. The current Polish Water Law is outdated, therefore there is a need to replace it with a law that meets current needs. The process of elaboration of the new Water Law is taking longer than expected (implementation was promised for 1 January 1995). The present situation brings many negative consequences (e.g., costs related to the function of Catchment Councils, which are already prepared to take over their new duties). Therefore, it is necessary to apply the new system as soon as possible.

A system like this one might become a base for constructing a complex system of protection. Adjusting administration to a regional level across the country will make it possible to modify instruments and standards for local needs. It will ensure a better realization of sustainable development in regions where water is short. Later, it will be possible to consider an introduction of additional instruments for the protection of water resources, the core of the system should remain at its proposed shape (with possible later corrections related to observation of its function). It is important to introduce a continuous evaluation, checking, and improvement of effectiveness concerning the water management policy. The system should follow a cycle of introduction of corrections, observation of effects, and introduction of further corrections.

References

Environment Law, 1980, Protection and formation of environment from day 31 of January 1980 (together with later changes), official gazette announcing current legislation, No. 49, Warsaw, Poland (in Polish).

Kleczkowski, A.S., *et al.*, 1984, Evaluation of water management, in St. Kozłowski, ed., *Management of Natural Resources*, PWE, Warsaw, Poland (in Polish).

Koza, I., 1994, Foundation system of the new water law, in *Protection and Exploitation of Water Resources*, Seminar Summary, Toruń, April 14–15 (in Polish).

Kozłowski, St., 1991, *Economy and Natural Environment*, PWN, Warsaw, Poland (in Polish).

Łustacz, E., 1991, Economic interpretation of water management, in *Social-Economic Conditions of Water Management*, Wydawnictwa Politechniki Warszawskiej, Warsaw, Poland (in Polish).

Pakulska, J., 1998a, Regionalization of water management in Poland, in *Market and Social Development*, Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin, Poland (in Polish).

Pakulska, J., 1998b, Payments and ecological penalties as economic tools to begin sustainable development in water management, in *Mechanisms for Conditioning Sustainable Development*, Conference, Rajgród, September 21–24, Politechnika Bialostocka, Bialystok, Poland (in Polish).

Pakulska, J., 1996, Water Management in Poland and Switzerland, typescript, IRG SGH, Warsaw, Poland (in Polish).

Rule, 1991, On the matters of classifying waters and of the quality of sewage introduced to waters or to the soil from day 5 of November 1991, official gazette announcing current legislation, No. 116, Warsaw, Poland (in Polish).

Symonowicz, A., 1988, *Use and Protection of the Environment*, SGPiS, Warsaw, Poland (in Polish).

Water Law, unpublished proposal, typescript (in Polish).

Chapter 19

Model of Municipal Water Pricing

Paweł Bartoszczuk

Abstract

This chapter describes the model of price setting, illustrating the concept with special reference to municipal water supply. The various elements influencing the price of water – namely, production of water, sale of water, and consumption of water – are described and a numerical example is provided. The following hypothesis is formulated and discussed: decrease of water consumption in households leads to a significant increase of municipal water prices. The chapter also reviews some consequences of decrease of water consumption: increase of water pricing, decrease of the amount of sewage dropped to the treatment plant, and higher cost of wastewater treatment. The concluding part proposes one of the types of water tariffs in waterworks. While the focus of the chapter is on estimation and implementation of the model applied to municipal water supply, the approach and techniques employed are of general relevance to natural monopoly.

Keywords: Cost of water production, natural monopoly, price, supply, water consumption, water resource.

19.1 Introduction

Establishing a reasonable price for water is very important both for the waterworks and consumers. The price should be based on the cost of supply, which means that it includes the operation, maintenance, and capital costs. The value of water has to be directly related to the costs of its provision so that a more rational use of water may be reached. The price charged for water should also depend on the location of the town. The process of water production requires large capital investments and has serious health implications for the community.

Water service is considered a natural monopoly because it is inefficient to have more than one service provider in a given geographic area. Budgets for carrying out its processes are constrained by the need to set a low price on a necessity of life. Water price has a significant impact upon water consumption, and therefore should reduce wasteful and inefficient use and improve resource allocation. If the price is too low, adequate services could not be provided. Resistance to price increases for water is often based on the argument that the poor must have access to supplies sufficient to meet their basic health needs. The real economic cost of water is usually a very small fraction of most households' disposable income, and so this argument is rarely true. Nevertheless, social and political reasons may require that water prices for a smaller consumer should be subsidized.

On the other hand, too high a water price hurts the consumer, particularly those industries that use large amounts of water or low-income consumers. Higher prices lead to lower consumption, especially in the short run when incomes are fixed. Consumption falls as people conserve water to avoid higher bills. Leaks are repaired and water use is reduced. Large consumers relocate or introduce water saving measures that considerably reduce the use of water. One of the factors influencing water consumption is installation of water meters. Households that are metered use up to 50% less water than those that have no meters, even without price increases.[1]

In the previous period of command economy, excessive consumption of water was observed in Eastern Europe. Water was relatively cheap in relation to other goods in that part of the world. The then-existing prognoses assumed the increase of consumption due to a forecast higher demand of the water-consuming industry. This trend was inverted with the application of market economy rules, when the price depended more on the cost of water supply. A considerable number of industrial enterprises decreased water consumption, and it became clear that the existing waterworks had excess capacity. One of the decisive factors in water consumption reduction was installation of water meters in households. Additionally, reduction of water losses induced a significant decrease in water consumption by households.

[1]Environment Waterworks Environment Canada, http://www.ec.gc.ca/water/en/info/pubs/ntttw/w_nttwi8.html

The objective of this work is to present and apply a method of establishing water prices. This method allows us to assess whether the process of increase of drinking water prices will be convergent or divergent. To realize this goal, relations will be presented among factors influencing water prices: the average cost of water treatment in municipalities, and production, and sale of water by the waterworks companies.

In countries like Poland, water price is set on the basis of planned annual costs divided by planned sales of water. This method of setting price does not take into consideration the elasticity of water demand. The price for industry is very often equal to or even higher than the price for households. Thus industry must pay more, even though the cost of service to industry is lower than that to households. The possibility that industry would considerably reduce the use of water was not taken into account by decision makers setting the price. Pricing for households is set by a political decision and often is too low to cover operational costs. It leads to a permanent lack of financial assets to cover costs, resulting in poor water services. Such policy leads to a decrease of water consumption.

A new approach presented in this chapter allows one to simulate the process of water price increase and decrease of water consumption. This model can be implemented in waterworks to assess whether a decrease of water consumption caused by higher prices will be serious and if it leads to lower utilization of waterworks capacity. By changing the relation between the price of water and average cost, a company can lead to a better utilization of waterworks capacity. It is very important, particularly in relation to large consumers. If they have to pay higher prices, thus subsidizing households, they will relocate or introduce water saving measures, considerably reducing the use of water. In conclusion, some implications of decreasing water consumption in households will be presented.

19.2 The Hypothesis

The results of this chapter allow formulation of the hypothesis that the decrease of water consumption in households leads to a significant increase of water price. Water saving in households leads to a decrease of water production by waterworks and declining utilization of the waterworks capacity. Consequently, the decrease of water production induces the growth of the average cost of water production. In the waterworks the total cost consists of the total variable cost and total fixed cost:

$$TC = TVC + TFC \, , \tag{19.1}$$

where TC is the total cost; TVC is the total variable cost; and TFC is the total fixed cost (Dziembowski, 1983).

Water delivery service has a high fixed cost and a low variable cost. The total fixed cost (TFC) amounts to as much as 80–90% of the total cost, and the rest is the total variable cost (TVC), which depends on production of water (Moll, 1995). In addition, the overprojected capacity of waterworks, which is due to overestimated consumption, is common in the former communist countries and it contributes to the higher percentage of fixed costs. Because of the high share of the total fixed cost in the total cost, the total cost strongly depends on the fixed cost. By dividing both sides of Equation (19.1) by water production (P) we obtain the average cost:

$$AC = AVC + \frac{TFC}{P} \,, \tag{19.2}$$

where AC is the average cost; AVC is the average variable cost; and TFC is the total fixed cost.

While the average variable cost (AVC) is constant, the average cost (AC) depends on production of water. A decreasing cost of water production gives rise to natural monopoly. A larger output means lower average costs per unit, and only one firm can survive on the market. If there were two suppliers on the market, one of them could expand to reduce costs and thereby eliminate the other. This situation precipitates a pricing problem because the surviving producer may be able to set prices well above the prices that would rule under competitive conditions. This is often the argument for regulating or nationalizing a natural monopoly.[2]

The utilities are regulated by the government, to prohibit them from exploiting their monopolies with higher prices. Some municipalities and government districts own the local utility and provide the service as non-profit. In such a case the municipality is a regulator, which acts as a substitute for the market, taking on some of the functions of competitors and attempting to provide similar incentives to improve efficiency by regulating aspects of the firm's conduct. Typically, the suppliers are allowed a low, fixed percentage of profit above cost.[3]

A monopolist is driven by the local council to price at average cost, thus enabling the firm to just cover its costs.[4] The price of water delivered to households is set on the basis of that total average cost. After some period of time, when the cost is growing, price of water must grow accordingly. This leads again to a drop of water consumption, causing an increase of the average cost of water under conditions that further lower utilization of the waterworks capacity which next leads to an increase of water price. This is particularly evident in countries where the price is inversely proportional to water consumption.

---

[2] Dnes, A.W., Franchising and Privatisation, World Bank, http://www.worldbank.org/html/fpd/privatesector/ppi-pubs.html

[3] Foldvary, E.F., Natural Monopolies, Progressive Report, http://www.progress.org/archive/foldvary.html

[4] Natural Monopoly, http://wiliam-king.www.drexel.edu/top/prin/txt/Monch/mon25.html

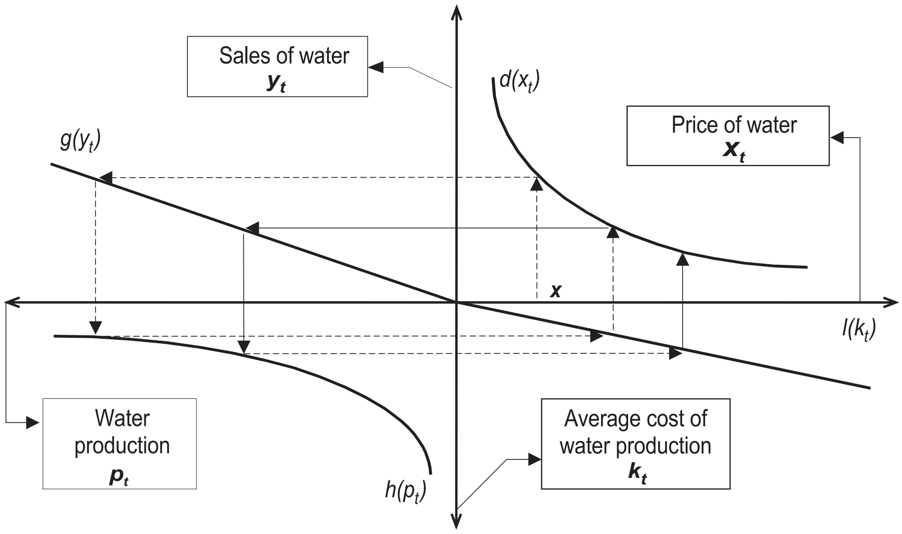


Figure 19.1. Graphical presentation of model relationships (19.3)–(19.6) among sales, production, average cost, and water price.

In the discussion of this hypothesis, a mathematical model was applied for analysis of the process of generating successive values of price of water. This concept applies the criteria of convergence of the sequence of successive approximations of solutions of recurrent model and also the theory of recurrent sequences (particularly conditions of convergence of these sequences).

19.3 Presentation of Relations Applied in the Model

For the presentation of the model the following functions were needed: the sales of water as a function of price, water production as a function of sales, average cost of water production as a function of production, and price for the 1 cubic meter (m^3) of water in the period $t + 1$ as a function of cost in the previous period (see *Figure 19.1*), where the following variables are used for each period indexed by t:

y_t = sales of water
p_t = water production
k_t = average cost of water production
x_t = price for the 1 m^3 of water in the period t

Equations (19.3) through (19.6) show the model relationships.

$$y_t = d(x_t) \tag{19.3}$$

$$p_t = g(y_t) \tag{19.4}$$

$$k_t = h(p_t) \tag{19.5}$$

$$x_{t+1} = l(k_t) = l(h(p(g(d(x_t))))) \tag{19.6}$$

The variable "sales of water" depends on the water consumption. The water consumption in a community is characterized by several types of demand, including domestic, public, commercial, and industrial uses. Domestic demand includes drinking water, and water for cooking, washing, and laundering. There is a wide variation in the total water demand. This variation depends on population, geographic location, climate, and the extent of industrial activity.

Demand for a product generally reacts to a change in price. When the price goes up, demand goes down accordingly. Elasticity of demand is a measure of how much demand changes in response to a change in price. When water meters are applied, the theoretical price-demand relationship applies:

$$Q = kP^e \ , \tag{19.7}$$

where Q is the demand for water; P is the price per unit of consumption; k is a constant; and e is a coefficient which measures the elasticity of the demand.

Since price increase will tend to depress demand, Q is proportional to the inverse of P, hence e is negative. When e equals 0, changes of prices have no effect on demand (Crowley *et al.*, 1994). There is a big inelasticity of demand in households: When $e = -1$ then Q is proportional to $1/P$ (reverse of P); small changes of P cause almost proportionate changes in Q. A low e value indicates a high degree of inelasticity. When e equals –0.2, a 29% price increase is required to reduce demand by 5%. The drop in unitary water consumption is a more serious problem in small towns than in bigger ones. A drop in water consumption of 10% causes a 70% increase in water prices, when elasticity equals –0.2. However, when elasticity equals –0.5, there will be only a 23% increase in water prices.

In a number of developed countries – Israel, Canada, the United States, Australia, and Great Britain – empirical analysis has shown that the price elasticity of demand for water in households is between –0.3 and –0.7, (i.e., a doubling of the price of water would reduce consumption by between 30 and 70 percent.[5] A similar range of elasticities is reported in studies of a number of developing countries in Asia and Latin America. There is also much empirical evidence about the potential substitution of capital for water in industry. For example, substantial increases in industrial water prices in Japan in the 1970s stimulated major investments in recycling, and sharp reductions in consumptive water use.

On the basis of research at 211 waterworks in Poland, average consumption in those waterworks in 1998 was 129.8 liters per capita per day (l/cap/day) (Bylka, 1998). However, this indicator for the entire country was close to 140 liters. For the individual water companies, average consumption ranges from 45 l/cap/day to 230 l/cap/day. In six waterworks it was more than 200 l/cap/day and in four lower than 50 l/cap/day. By assuming that the cost of producing one cubic meter of water is the same, the water price in the town using 200 l/cap/day should be considerably lower than in a town using only 50 l/cap/day. In fact, the difference in 1999 water prices in towns where the consumption was lowest and highest was not bigger than 20% (Bylka, 1999). Conversely, the highest price was in towns where the indicator was very high. The changes in average water use in Poland in previous years are shown in *Figure 19.2*.

Current average indicators of water consumption in Poland are similar to German ones and lower than it was forecast in the command economy. The use of water in Poland now is at the average European level, or even less than in other European countries. In Western countries, e.g., in Austria and Italy, the use of water is higher than 200 l/cap/day. In Japan and Canada more than 250 l/cap/day of water is used, and in the USA consumption is at 380 l/cap/day.

However, most people in the world (living in Africa and Asia) use only 50 l/cap/day, which is less than what they need.[6] The World Health Organization has established a minimum of 150 liters per capita per household in the cities of the developing world; 75 liters per day per household are considered adequate to protect against waterborne diseases. The World Bank estimates that at least 50 l/cap/day are necessary for sanitation, in order to forestall health-related problems. Some evidence suggests that residents of poor settlements use as little as 20 l/cap/day. This is sufficient for cooking and drinking, but not for maintaining a healthy environment. On the other hand, higher-income households typically consume many hundreds of l/cap/day for discretionary purposes such as lawns and pools.

[5]Warford, J.J, Marginal Opportunity Cost Pricing for Municipal Water Supply, http://eepsea.org/publications/spaper/Warford.html

[6]Environmental Works, Encyclopedia Britanica, http://www.britannica.com/article/7/0,5716,127657+12+117295,00,html

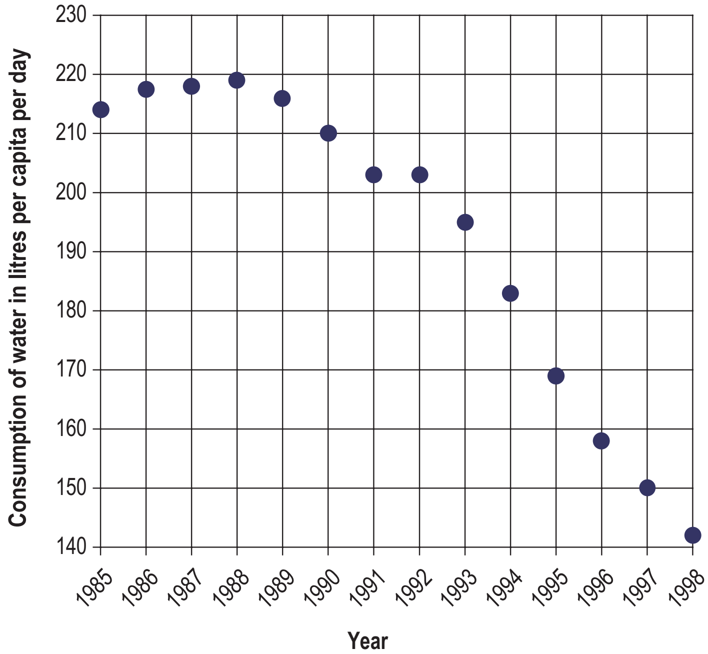


Figure 19.2. Consumption of water from municipal waterworks in Poland. Source: Water Resources in Poland (1996) and Polish Statistical Office (1998).

By eliminating the variables "sales of water," "production of water," and "average cost" (y_t, p_t, k_t) from the model Equations (19.3)–(19.6), a recurrent process was obtained: water price in a period $t+1$ in a function of price in the period t:

$$x_{t+1} = f(x_t) \,, \tag{19.8}$$

where $t \in \{1, 2, 3 \ldots\}$.

Let us consider that $x_0 \in R_t$ belongs to the domain of function $f(x)$, that is, by assuming water price in a current year (set on the basis of average cost of water production). By substituting $t = 1$ for price x_1 we get the price in the next period $f(x_1) = x_2$. Then for $t = 2$ we receive price in the next period $f(x_2) = x_3$, and so on until we receive the price x^*. The sequence $x_{t+1} = f(x_t)$ is convergent to that price and its boundary is the value x^* (see *Figure 19.3*). The price is set on the basis of planned annual cost of water production. If the real consumption of water in the next period is lower than planned, the waterworks will have financial losses. Price x_2 is certainly higher than price x_1, because an increase in the cost of water production is caused mainly by a decrease of water production. But in some

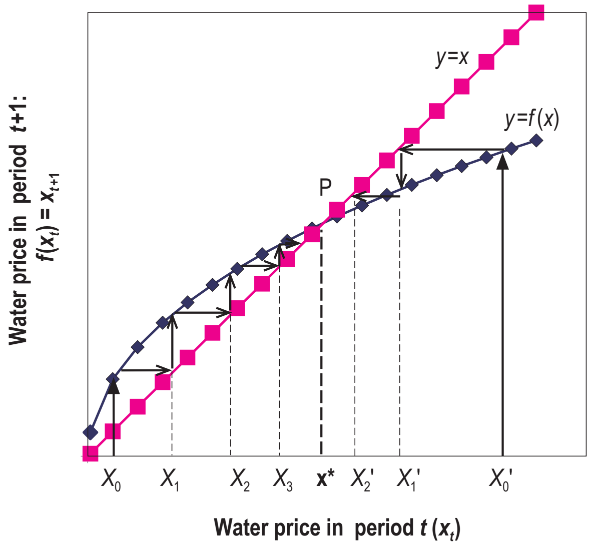


Figure 19.3. Process of price creation: scheme with an assumed shape of curve $f(x)$.

cases, price setting is a political decision and the water price could be higher than previously, while not covering the average cost. In such cases, municipal authorities must subsidize the waterworks.

On the other hand, the increase in water prices depends on water policy, i.e., the relationship between the price of water and water costs. It was assumed that prices are considerably higher than the costs of water supply. By setting the price at a level equal to or lower than the cost of water production, the waterworks can stop the process of decreasing water consumption. Lower price of water leads to a greater water sale and therefore to a larger utilization of the capacity of waterworks (Suligowski, 1997). As can be seen in *Figure 19.3*, the price increases in the next periods by smaller and smaller amounts. Decrease of water consumption is associated with increase of water price (Poss and Hacker, 1991; Bagieński and Stodulski, 1996). It can be concluded that decrease of water production is lower in the successive periods. The possibility of decrease of water consumption will be finished. It means there is a minimum level of consumption of water, by which the price will be stabilized (see *Figure 19.3*).

Convergence of the sequence does not depend on the choice of the starting point x_0. By substituting for x_0 any positive real number another sequence of successive approximations could be obtained which is always convergent to the boundary x^*. When $x_0 < x^*$, then the successive approximations form an increasing sequence (assuming $f(x)$ is an increasing function), so the values of this sequence will increase until the value of point P. Otherwise (i.e., for $x^* > x_0$), it is a decreasing sequence, bounded from below by the abscissa of the point P.

19.4 An Example of Application of the Model

The parameters for Equations (19.8)–(19.11) were obtained from the data collected from the waterworks companies in Polish towns in the 1990s. It was assumed that the relations between variables of the model were constant in the long run.

$$\text{Sales of water}: \; y = 78.829^* x_t^{-0.59} \tag{19.9}$$

$$\text{Water production}: \; p = 1.133^* y + 1.884 \tag{19.10}$$

$$\text{Cost of water production}: \; k = 699.3^* p^{-1.2412} \tag{19.11}$$

$$\text{Water price per the } 1\text{m}^3: \; x = 1.599^* k^{0.70} \tag{19.12}$$

The relation between sales and water price was approximated by the power line (19.9) [the same applies for relations (19.11) and (19.12)]. Relation (19.10) was approximated by a straight line.

By eliminating the variables "sales of water," "production of water," and "average cost" (yt, pt, kt) from the model Equations (19.9)–(19.12), the recurrent relation was obtained:

$$x_{t+1} = 3.1808^* x_t^{0.4806}\,, \tag{19.13}$$

where x_{t+1} is the price for the period $t + 1$, as a function of price in the period t.

For example, for an initial price [4.86 zloty per cubic meter (zl/m^3)] set to the actual price in one of the towns, the recurrent process (19.13) with this initial price is convergent to the price x^* = 9.27 zl/m^3, which is an abscissa of the point, in which the line $y = x$ crosses curve $F(x)$. The number x^* is a boundary of sequence (19.13). For this price we obtain the sale of 21.2 million m^3 of water per year, which is identical to two Polish towns: Gorzow and Kalisz. From these calculations one observes that decreasing water sales cause a significant increase in water price. The calculations are valid for a domain for which the model is still valid. In reality, the authority will probably not allow the water price to get too high and will have to provide subsidies.

19.5 Conclusions

The results of the studies show that a decrease in water consumption leads to an increase in water prices, when there is no fixed part of water tariffs. It follows from the presented method that there is a possibility that the process of decrease of water consumption and increase of water price will end. It is not clear whether this will happen, because inhabitants of towns in Poland and some other countries in Eastern Europe continue to decrease their consumption.

Unfortunately, the law in some countries – including Poland – prohibits the application of the fixed part of the price. The following approach could be applied in practice. One could use a tariff schedule for households that consists, for example, of two steps: a low subsidized rate for the first 5 to 10 cubic meters per month and a higher rate for all additional consumption. If the second step equals marginal cost, it may result in an acceptable trade-off between economic efficiency on the one hand and equity (lower price for a necessity of life) on the other. Such an incentive-oriented pricing structure would be more relevant for economic and environmental issues of water management.

Several consequences arise from decreasing water consumption. First, saving water is good because less of the environmental resources are used. Second, the higher price of water and probably the cost of processing sewage are not welcome by the households. Third, the decreasing water sales lead to a decrease in the amount of sewage dumped to the sewage treatment plants. This induces a higher concentration of impurities in the sewage when the load of impurities is constant. Such a situation can cause big exploitation problems at the wastewater treatment plants. It can require changes in the biological part of the existing plant. Such investments can be very expensive. If there are not sufficient financial means to support the change of technology, we can even pollute the environment by incompletely cleaned wastewater. This is a paradox: an ecologically minded activity can lead to pollution.

On the other hand, the decrease in the amount of water and sewage can lead to a new situation, when both new waterworks and a new sewage plant can be smaller and cheaper. To assess further consequences of the decrease of water consumption, particularly with respect to the cost of sewage treatment, further studies are required.

References

Bagieński, J., and Stodulski, J., 1996, Wpływ Instalacji Wodomierzy Mieszkaniowych na Zużycie Wody w Budynkach (Influence of Installation of Watermeters in Households on the Consumption of Water), Zaopatrzenie Miast i Wsi w Wodę (Water Supply), conference papers, Poznań, Poland.

Bylka, H., 1998, Kształtowanie opłat za wodę i ścieki. Możliwości i ograniczenia (Water and sewage price setting. Possibilities and limitations), *Przegląd Komunalny*, **3**.

Bylka, H., 1999, Kilka uwag o stanie wodociagów i kanalizacji (Remarks about Polish waterworks), *Przegląd Komunalny*, **4**.

Crowley, F., Law, F., Twort, A., and Ratnayaka, D., 1994, *Water Supply*, Edward Arnold, Kent, UK.

Dziembowski, Z., 1983, *Ekonomika przedsiębiorstwa komunalnego (Economics of Communal Enterprises)*, Państwowe Wydawnictwa Ekonomiczne, Warsaw, Poland.

Moll, H.G., 1995, Wassersparen um jeden Preis? Nein Danke!, *Neue DELIWA Zeitschrift*, **8**.

Poss, C., and Hacker, G., 1991, Maximale Tages- und Studenabgabe an Versorgungsgebiete in der Trinkwasserversorgung der Bundesrepublik Deutschland, *GWF Wasser/Abwasser*, **11**.

Stan i Wykorzystanie Zasobów Wód Powierzchniowych Polski, (Water Resources in Poland 1996), 1996, Instytut Meteorologii i Gospodarki Wodnej, Warsaw, Poland.

Suligowski, Z., 1997, Oszczędzanie Wody (Water saving), *Gaz Woda i Technika Sanitarna*, **1**.

Part V: Applications

Chapter 20

Bioenergy as a Factor in the Sustainable Development of Rural Areas in Poland

Wiesław Ciechanowicz

Abstract

In this chapter we suggest ways to address rural area development problems via the bioenergy to be used in the transport sector. The next decades will bring an increasing demand for methanol, according to information given in the literature. The Polish regional bioenergy companies, together with foreign firms, could become substantial suppliers not only of world methanol, but of the fuel cells that power buses. The System Research Institute and the Bug River Association Authority agree that biomass and bioenergy are the major factors leading to sustainable regional development.

Keywords: Sustainable rural area development, biomass, bioenergy, automobile powered by direct methanol fuel cells.

20.1 Introduction

One of the Polish economy's main problems is rural area development. About 12% of the state budget is transferred each year to rural areas as a supplementary subsidy to agricultural products and in other forms of subsidies. About 12% of inhabitants are unemployed, 75% of whom are uneducated, and most of them live in rural areas. Given the following circumstances, it seems that biomass cultivation should be one of the solutions to the sustainable development of rural areas:

1. Biomass can play a prominent role in a restructured marketplace, it should, in particular, replace rye and potatoes as the main plants cultivated on poor arable land in Poland, provided the biomass production costs are competitive with those of fossil fuels.
2. Since biomass is bulky and costly to transport, biomass conversion facilities must be situated close to where the crop is grown [not farther than 30 kilometers (km)]. This requirement would mean that products of rural areas would be sold as final products rather than as raw materials, which is the current situation. In this way new jobs would be created.
3. The rapid increase in the concentration of greenhouse gases, especially carbon dioxide, in the atmosphere threatens global climate changes, resulting mainly from the large-scale use of fossil fuels. Biomass, if produced in a sustainable manner, can be a carbon-neutral alternative to fossil fuels.
4. Not only can biomass energy help reduce waste and clean up the environment, it can also provide new and better jobs for rural Poland.
5. The environmental benefits of biomass compared with the negative impacts of fossil fuels should allow biomass to compete economically.

Despite the promising outlook, a wide range of barriers is hampering the large-scale development and implementation of commercial biomass energy systems. First, the specially cultivated biomass is generally too expensive. However, biomass is able to compete on a significant scale in countries like the USA, Sweden, Denmark, and Brazil, where government policies support its use financially or have actively discouraged the use of fossil fuels (e.g., by introducing the carbon tax).

Second, as mentioned, about 12% of the state budget is transferred annually to the rural areas as a subsidy to agricultural products in Poland. This can be viewed as too expensive for the economy. Third, biomass has a relatively low energy density. Fourth, the complexity of large-scale bioenergy systems is also a barrier. Production of biomass is bound up with seasons and makes high demands on organization and logistics. Moreover, it should involve many different actors connected with the production and use of energy crops: e.g., farmers, utilities, industries, agriculture and environment protection ministries, and government. Fifth, difficulties with public acceptability and uncertainties concerning long-term investment in the large-scale production and use of biomass form another problem.

Fundamentally, two possible scenarios exist for rural area development in Poland:

1. Building a sustainable rural area development based on bioenergy, or
2. Doing nothing but still subsidizing agricultural product prices and spending quite large amounts of money on allowances for the jobless in rural areas.

This chapter has the following specific objectives:

- analysis of the current situation, and
- presentation of the major factors that may facilitate a sustainable development in rural areas.

20.2 Key Issues Involved with the Agricultural Sector

20.2.1 Analysis of the situation

For Poland, in addition to solving some environmental protection problems, sustainable development means a decrease in the economic differences between rural and urban areas. The vision of rural area development proposed here should in the long term realize similar living conditions in both rural areas and urbanized regions.

The main problem of long-term development in Poland is the development of the primarily rural regions, including agriculture. The shorthand information in *Table 20.1* can to some extent justify this statement (Statistical Yearbook of the Republic of Poland, 1998).

20.2.2 The scale of the required capital

Assuming that, due to farm restructuring, agricultural employment would decrease to 3% of total population, and taking into account the forecast of the demographic growth along with the current unemployment of 0.65 million in these areas, the expected number of the unemployed in rural areas could come to 5.5 million. Thus, taking a longer perspective, the development of rural areas cannot be achieved by

Table 20.1. Information on Polish agriculture.

- Poland is the country with the vast majority of rural-type communes.
- The agricultural sector, which employs as many people as does industry, has a share in the GDP which is five times smaller.
- Employment in agriculture was approximately 12.9% of the country's population in 1996.
- The soils of about 60% of arable land in Poland are poor, and 30% is very poor.
- Farms with arable land area below 7 ha account for 68% of the total number of farms.
- There is a potential for doubling the plant yield in Poland, given the levels observed in the EU.

Source: From Statistical Yearbook of the Republic of Poland (1998).

investing in the agricultural sector only. Other economic activities, such as manufacturing, housing, tourism, and services will have to also contribute to the development of rural communes.

According to an initial assessment (Ciechanowicz *et al*., 1999), the capital required for achieving sustainable development of rural areas might even come to US$160 billion, including job creation for 5.5 million people, education, and infrastructure. The current budget of the Polish economy now equals about US$40 billion (Ciechanowicz *et al*., 1999).

If, however, we assume that agricultural employment due to farm restructuring would merely decrease to 10.5% of the total population, given the forecast of demographic growth and the current unemployment in these areas, the capital required to create new work places for 2.5 million people would still amount to US$80 billion. That means the scale of the required capital would still be double Poland's budget. Thus, private capital will have to play the key role in solving rural problems.

One must remember that the effects of the development to be supported by such funds can be observed only after 20 or 30 years. Hence, a development strategy must account for the following challenges:

- an education accessible to all in rural areas;
- the simultaneous fighting of both water deficit and floods, by constructing small water retention systems;
- a fulfillment of stronger requirements for environmental protection by:
  - biomass cultivation as a short rotation intensive culture of woody biomass (Borjesson, 1998),
  - bioenergy utilization.

Responses to these challenges must be considered as major factors facilitating sustainable rural development.

20.3 Biomass as an Alternative for Rural Development

20.3.1 Attempts to initiate biomass cultivation in Poland

A computer system for regional development planning – accounting for energy and agriculture, including biomass, water management, environmental protection, and sustainable development of rural areas – was worked out in 1994 at the Systems Research Institute (SRI) with collaboration from the authorities of ten provinces. One of the questions discussed was how to simulate expansion of agriculture and of other activities in communes, these being the basic elements of the territorial autonomy in the model. This was implemented in the framework of the research project completed in 1998. Attempts were then made to attract the interest of the governmental authorities (Chief of Prime Ministry Council, Minister of Agriculture, and Minister of Finance) to support the sustainable rural development project based on the biomass cultivation with the suggestion of financing the problem.

Finally, the Association for the Bio-Energy Commercialization was formed by the Bug River Association, representing a dozen counties in the eastern part of Poland, the SRI, *Siedlce* Magazine, and the Ekoland Foundation.

Agriculture holds the key to bioenergy, since it supplies the utilities with the primary energy carrier – the biomass – to be converted into the secondary energy carriers. The waste wood from forests and the waste straw from agriculture cannot be considered to be the main priority for rural development. They will not create a true market for agricultural products nor new jobs. Therefore, the main interest would be in biomass as a true crop produced for energy purposes.

The study of the potential for bioenergy in Poland through dedicated energy plantations was undertaken at the SRI. The aim of this study was to explain questions regarding the advantages of biomass cultivation, taking into account the following aspects:

1. The biomass available in communes, counties, and provinces.
2. The incomes of farmers.
3. The numbers of new jobs created.
4. The incomes of communes as a share of taxes.
5. The government's gain due to the decrease of subsidies to agriculture.

20.3.2 The available biomass energy potential

Let us evaluate the potential energy biomass resource for Polish economy. Willows can be planted on soils of the worst quality, i.e., on 6 million hectares (ha) out of 18 million ha of arable land in Poland. Assume that these 6×10^6 ha will be used for Short Rotation of Intensive Culture (SRIC). Let us consider two possible wood biomass yields per hectare, namely 20 and 30 dry t/ha/year. Then the potential productivity of wood biomass for NCV of 20 gigajoules/ton (GJ/t) would be equivalent to 96–144 tons of coal with NCV of 25 GJ/t. That means that virtually all collieries in Poland could be closed within 30 years. As a consequence of creating new jobs in rural areas, miners would have lost their jobs. Therefore, we assume that the main aim of biomass cultivation is to replace the oil import rather than the coal fueled power plants in Poland.

This estimation of available biomass potential is of course quite rough. A more accurate assessment of short rotation woody crop production would require consideration of biological and operational issues associated with growing woody crops in managed plantations. Farming policy, infrastructure support, financial arrangements, conversion technology requirements, potential risk, and local environmental conditions should be considered as well in order to make informed decisions about location and management. Successful ventures also require regional or local crop development and research activities to assure the availability of selected materials that are adapted to soil and climates of the region and the availability of knowledge and techniques for crop management. Different forms of co-operation with foreign companies would also be required.

20.4 Bioenergy

20.4.1 The main purpose of bioenergy usage in rural areas

The reason for investigating bioenergy utilization is the improvement of the living standards in rural areas. There are two possible applications of biomass:

1. Stationary power (Bajura, 1999; Hirschenhofer *et al*., 1998; Wiens, 1999) or
2. Vehicle motive power (Wiens, 1999; Zalbowitz, 1998).

The technologies for the first option could be the biomass gasifier/integrated gas turbine combined cycle power plant or the biomass gasifier and Solid Oxide Fuel Cell (SOFC) as the combustion chamber of a gas turbine.

For the second option the biomass converter to methanol and Direct Methanol Fuel Cell (DMFC) technology are required.

In the first case the biomass would replace coal and in the second case, imported oil. The possibility of eliminating coal mining and coal-fired plants in Poland in the time horizon of 20–30 years is doubtful. A powerful coal lobby and very strong miners' trade unions exist. While improving their living standards, the farmers would lower the living standard of the people of Upper Silesia. Hence, the compromise seems to be vehicle motive power, particularly as a source for city automobile power.

The SOFC could be considered for the export purpose in the future as a result of the possible collaboration of comparatively low-labor-cost regions in Poland with foreign companies involved in research. Such collaboration would create new jobs in these regions. One should mention that the co-generation aspect of SOFC would be extremely important for countries with colder climates. Fuel cell power plants installed at an apartment complex could provide the electricity for the building as well as heat and hot water for the building's tenants.

20.4.2 Fuel cells

In the last few years research has made critical strides in developing commercially viable fuel cells, namely the SOFC that can extract electricity from natural gas, ethane, and other biomass and fossil fuels, and the DMFC, which can be fueled directly by methanol.

Fuel cells convert chemical energy of fuels directly into electricity. The principle of the fuel cell was developed by William Grove in 1839. Around 1900, scientists and engineers were predicting that fuel cells would be common for producing electricity and motive power within a few years. That was 100 years ago. Contrast this with the roughly two years it took Nikolaus Otto to bring his 4-stroke internal combustion engine from the invention stage to a commercial success.

Fuel cells are being proposed to replace the Otto or Diesel engines as reliable, simple, quieter, less polluting, and of an even greater economy, provided they are of the DMFC type. However, this type of fuel cell appears also to be the most promising technology for small power plants and as battery replacement.

Several companies around the world are presently working on DMFC. In 1999 there was a market shift away from developing polymer electrolyte fuel cells (PEFC) in favor of the DMFC. In this type of fuel cell, methanol is not reformed into hydrogen gas but is used directly. Tremendous progress has been made in the last six years. The efficiencies of the DMFC are much higher than in the early 1990s, and predicted efficiencies in the future may be as high as 40% for the Diesel cycle automobile power plant. Moreover, it is expected that the DMFC will be more efficient than the PEFC for automobiles that use methanol as fuel.

If vehicles rather used hydrogen as fuel, a hydrogen supply system would need to be installed. This would be extremely expensive. The DMFC, however, would likely be simpler than the internal combustion engine, producing superior efficiency and being less polluting. The liquid fuel could be handled by slightly modifying the present distribution equipment. When the DMFC is perfected there may be a major swing away from the Otto and Diesel cycle automobiles. There are many vehicle manufacturers betting on this. Even Ballard, who has been the major proponent of hydrogen economy and the hydrogen fuel cell, bought major DMFC technology in August 1999 (Chubb, 2000).

Working in collaboration with Ballard, Daimler-Benz built a series of PEFC powered vehicles. Other major automobile manufacturers, including General Motors, Volkswagen, Volvo, Honda, Chrysler, Nissan, and Ford, also have announced plans to build prototype fuel cell vehicles operating on hydrogen, methanol, or gasoline.

A few buses fitted with fuel cells are already on the road. Ballard expects that by 2003 there will be 55 such vehicles on the streets of California. In 1997, Daimler Chrysler developed Nebus, a demonstration bus powered by the fuel cell. Collaboration between a Norwegian project and Daimler Chrysler was established in 1998 to focus on this fuel cell technology. In August 1999 in Oslo, a new hydrogen fuel cell bus, also called "Nebus," was tested for the first time under normal public transport conditions. It showed good driving performance, particularly in dense city areas, and the prospects for the technology are promising.

As far as cars are concerned, it looks as if the first company to enter into commercial scale production will be Daimler Chrysler, which expects to have a version of its "Necar" on sale in 2004, right on time to comply with the environmental legislation. The latest fuel cell system to be shown to the public by Daimler Chrysler is the Ballard Mark 900, which delivers 75 kilowatts (kW) on methanol and 80 kW on hydrogen. It measures roughly one meter in length by 30 centimeters (cm) in height by 45 cm in width. This gives it a volume of roughly 77 liters, which easily fits into a standard engine compartment of today. It supplies power at temperatures as low as 25°C. Two days after Daimler Chrysler introduced the Mark 900, Ford unveiled a prototype, a "Ford FC5," which also used a new fuel cell technology.

With respect to use of fuel cells in stationary energy power systems, Siemens Westinghouse focuses on SOFC commercialization, offering a hybrid fuel cell/gas turbine plant with an expected efficiency of $\sim$70%. The fuel cell module replaces the combustion chamber of the gas turbine engine. Two versions are considered: the coal fueled SOFC and the natural gas fueled pressurized SOFC.

The studies conducted in the USA (Bajura, 1999) suggest that by 2010 fuel cells could account for as much as one-tenth of the US$50 billion per year global market for power generation equipment. SOFCs in three sizes – 20, 100, and 250 kW – would satisfy most of the market capacity requirements. Moreover, the goals for the 21st century fuel cell program are to develop advanced solid state fuel cell for hybrid systems, smaller than 20 megawatts (MW):

- by 2010 with the potential for commercialization, an efficiency of 70% and an installed cost of US$1,000/kW;
- in 2015 with installed costs approaching US$400/kW and efficiency ~80% (Bajura, 1999).

In the framework of the Federal Energy Technology Center, USA, the fuel cell research budget for molten carbonate fuel cells, solid oxide fuel cells, and hybrid fuel cell/gas turbine systems is at around US$50 million per year.

20.4.3 Methanol as a fuel for automobiles or as a product for export

Poland has the capacity to start a collaboration with foreign companies to produce automobiles that use methanol. However, even if this were not implemented, Poland could become a large exporter of methanol based on biomass conversion in the future. Certainly, there will be a demand for environmentally friendly fuel.

As mentioned, about 60% of arable soils in Poland are poor (10 million ha) and 30% are very poor. Let us assume the possible wood biomass yield/ha equals 30 dry t/ha/year. Then, the potential productivity of wood biomass for NCV 20 GJ/t would be equivalent to 240 million tons of coal with NCV 25 GJ/t.

20.5 Final Remarks

This chapter suggests that one of the realistic solutions to Poland's development problems in rural areas may be based on the bioenergy to be used in the transport sector. The following reasons justify this suggestion:

1. There are two possible applications of biomass: stationary energy power or vehicle motive power. There is no possibility of eliminating coal mining and coal fired plants from Poland in the time horizon of 20–30 years. Hence, the compromise seems to be the vehicle motive power particularly as a source for city automobile power.
2. In the last few years, research made critical strides in the commercial development of two types of fuel cells; namely, the SOFC and the DMFC, which can be fueled directly by methanol.
3. DMFC appears to be the most promising for portable applications like transport.
4. A small number of buses fitted with fuel cells are already on roads. The latest fuel cell system to be shown to the public by Daimler Chrysler is the Ballard Mark 900, which delivers 75 kW on methanol and 80 kW on hydrogen. Two

days after Ballard introduced the Mark 900, Ford unveiled a prototype "Ford FC5," which uses a new fuel cell technology.

5. Studies by General Motors and Ford noted that fuel cell car engines could be built for about the same price as internal combustion engines. Daimler Chrysler pledged to have a viable, commercial fuel vehicle available in 2004.

This information indicates that in the next decades we can expect an increase in demand for methanol. Polish regional bioenergy companies, together with foreign firms, could become substantial suppliers to the world methanol market and to buses powered by fuel cells.

The SRI and the Bug River Association Authority agree that biomass and bioenergy are the major factors leading to sustainable regional development. However, there is also a need for such understanding by the government; development of energy technologies has always come from a partnership between the government and the private sector.

Moreover, in the transition from the centrally planned to the free market economy the principle "the less of government in economy, the better for economy" cannot be applied in the same degree to all types of activities or sectors. In particular, it does not apply when activities need a longer time period for transfer to the market, and also when the need arises for a change of mentality of people dealing with such activities.

We strongly require a significant reduction in the greenhouse effect. One of the activities leading to this reduction is an increase in the rate at which biomass plantations are established across the globe, preferably in the short term. Time is of the essence. Thus, there is an urgent need to educate decision makers and society in general about such activities. We must work assiduously to reduce the huge distances in both time and space between those who are supposed to act upon environmental issues and those who will benefit.

References

Bajura, R.A., 1999, Fuel cells: Simple solutions in a complicated world, in *NETL Proceedings of Joint Fuel Cell Technology Review Conference*, available on Web: http://www.fetc.doe.gov/publications/proceedings/99/99fuelcell/99fc.html

Borjesson, P., 1998, *Biomass in a Sustainable Energy System*, Lund University, Sweden.

Chubb, L., 2000, Fuel cells, the next generation: Lean, mean and clean, *Environmental News Network*, 15 January.

Ciechanowicz, W., *et al*., 1999, Problems of Economy, Energy, Water Management and Environment in the Simulation of the Sustainable Development of Regions with the Majority of Rural Areas, presented at "IIASA Days in Ukraine," an international meeting in Kiev, March, International Institute for Applied Systems Analysis, Laxenburg, Austria and The Ukrainian Academy of Sciences, Kiev, Ukraine.

Foran, B., and Mardon, Ch., 1999, Beyond 2025: Transitions to a Biomass-Alcohol Economy Using Ethanol and Methanol, Working paper series 99/07, December 1999, CSIRO Wildlife and Ecology, Canberra, Australia.

Hirschenhofer, J.H, Stauffer, D.B., Engelman, R.R., and Klett, M.G., 1998, Fuel Cell Handbook, 4th ed., DOE/FETC-99/1076, November, Department of Energy, DOE/FETC, Washington, DC, USA.

Pearson, W., 2000, Tree plantations: Australia's future fuel source?, *Environmental News Network*, 24 August.

Statistical Yearbook of the Republic of Poland, 1998, Central Statistical Office, Warsaw, Poland.

Wiens, B.I., 1999, The Future of Fuel Cells, 12 December, Ben Wiens Science Site, Internet, www.benwiens.com

Zalbowitz, T.S., 1998, Fuel Cell: Green Power, LA-UR-99-3231, Los Alamos National Laboratory, Los Alamos, NM, USA.

Chapter 21

Stochastic Projections of Persons, Jobs, and Residences in Australia

Richard Cumpston

Abstract

This chapter describes the development of regional projection systems, intended to provide 50–100 year projections for each of the 2,600 postal areas in Australia.

The demanding requirements have led to the adoption of methods very different from those normally used for regional population projections. Instead of analyzing cohorts, individuals are simulated, and their probabilities of birth, partnership, home purchase, employment, and death are randomly generated. Simulating individuals gives the flexibility to model social and economic variables at fine geographic scales.

The models may help project the environmental effects of particular changes. Examples discussed are the Internet, rising oil prices, rural industry changes, and salinity. Long-term projections at fine geographic scales may be needed to design environmental counter-measures and gain public acceptance.

Keywords: Stochastic projections, population, employment, housing, markets.

21.1 Introduction

This chapter describes the development of regional projection models, intended to provide 50–100 year projections of persons, jobs, and residences for each of the 2,600 postal areas in Australia. This is an ambitious project, as attempts to find any similar models now working anywhere in the world have been unsuccessful.

In 1997 Barney Foran, of Australia's Commonwealth and Scientific Industrial Research Organization (CSIRO), asked us to make long-term regional projections. CSIRO sometimes needs to answer questions such as "Where should a proposed sewage works, with a 100-year lifetime, be built?" Questions like this one involve a longer time frame, and a more regional focus, than the population projections normally available from the Australian Bureau of Statistics or the state planning authorities.

The work rests on the belief that employment is the crucial factor in deciding where most people live. Not only people, but jobs must be modeled. This necessity requires larger models and the use of unfamiliar modeling techniques. If the challenging problems can be solved, the resulting models may have a range of applications. It is fortunate that the work is privately funded, and not subject to specific tasks or deadlines.

Australia is sparsely populated, with 19 million people living mainly on the coasts of 7,730,619 square kilometers (km^2) of land. Some of the examples in the chapter are for Victoria, Australia's most densely populated state, which has 4.8 million people in 227,767 km^2. Australia has 2.5 people/km^2, compared with 97 in Austria, 232 in Germany and 336 in Japan. Australia's sprawling cities and wasteful use of land may be unfamiliar in Europe, but more common in the United States and Canada.

This chapter is organized as follows: Section 21.2 discusses projection methods and migration assumptions. Section 21.3 involves employment-driven models and findings for various industries. Section 21.4 concerns house, land, and farm prices. Section 21.5 discusses migration behavior. Section 21.6 examines past and present views on modeling and on computer needs and programs. Section 21.7 analyzes data sources and synthesis. Section 21.8 looks at the environmental effects of the Internet, oil price increases, changes in rural industries, and salinity issues. Section 21.9 presents the conclusions.

21.2 Projection Methods

21.2.1 Cohort population projections

Population projections are normally made by cohort methods, where groups of persons of known sex and age are projected forward a year at a time, allowing for births, deaths, and migration. The numbers of persons in each group are projected, but not their individual characteristics. The assumptions underlying these projections are sometimes randomly varied, giving "probabilistic" population projections (see Lutz *et al*., 1996).

21.2.2 Stochastic projections of persons and households

By contrast, the models described here project individual persons and households, and randomly simulate the events happening to each one in each time period. Some of the events simulated are births, deaths, partnership formation and dissolution, persons entering and leaving households, employment, migration, and home purchase. Probability distributions are assumed for each event, and in each simulation period random numbers are drawn to determine which events actually happen. Job seekers are plausibly matched with vacancies, home buyers are matched with dwellings for sale, and persons seeking partners are matched with other persons seeking partners.

21.2.3 Storage efficiency of individual projections

Orcutt *et al*. (1961, p. 287) noted that several million cross-classifications are required to represent different types of household, allowing only for age, sex, and marital status of the head of household, and the number, age, and sex of children. Each possible combination of variable values is one cross-classification. Adding other information, such as income, assets and region, makes the numbers of cross-classifications impractical. For example, 10 variables, each with 10 possible values, give 10 billion cross-classifications to be stored. If these 10 variables were to be stored for each of 20 million people, 200 million values would need to stored. The storage efficiency of individual projections, compared with cohort projections, increases with the number of variables.

21.2.4 Migration assumptions

Cohort population projections often assume constant numbers of migrants, or constant migration rates, between regions. For example, the Australian Bureau of Statistics (1998), in its population projections for each Australian state from 1997 to 2051, assumed constant net numbers of migrants into each state. But the poor results from attempts to fit multiplicative probability models to migration between Australian states and statistical subdivisions suggest that the models are inadequate reflections of the complex patterns of migration, and have little predictive value (Cumpston and Sarjeant, 1998). Stochastic projections allow migration to be simulated allowing for industry changes and individual circumstances.

Table 21.1. Distances traveled from home to work, school, or shops.

Distance from home	Work (%)	School/ college (%)	Shops (%)
Less than 500 m	3.1	11.6	7.9
500 m to less than 1 km	4.3	18.1	14.6
1 km to less than 2	6.6	17.1	20.7
2 km to less than 5	17.0	19.1	22.3
5 km to less than 10	19.5	11.7	7.5
10 km to less than 20	22.1	10.6	4.4
20 km or more	15.6	6.2	5.2
Not applicable	11.9	5.7	17.4
Total	100	100	100

21.3 Employment-Driven Models

21.3.1 Distances traveled from home to work, school or shops

The survey data in *Table 21.1* are from Victoria and the Australian Bureau of Statistics (1995). The median distance from home to work was about 8 kilometers (km), showing that most workers like to live close to their work (and even closer to schools and shops). Because of this close link between employment and residence, the geographic dispersion of different industries is very relevant to regional population projections. Note that 84% of persons used a car as their main method of travel to work, and only 8% used public transport.

21.3.2 Employment locations and transport costs (Victoria 1998)

Most of *Table 21.2* is derived from the numbers of workers employed in each of the 17 industry groups in Victoria, as recorded in the Business Register of the Australian Bureau of Statistics. The numbers analyzed are for 195 statistical local areas, varying in size from 680 to 110,000 persons. The average distance of workers from the capital was estimated as 58 km, using great circle distances from the centroid of each statistical local area. For comparison, the average distance from the capital of persons living in Victoria is about 60 km, and the distance of the most remote statistical local area from the capital is 471 km.

21.3.3 Transport costs

The last column of *Table 21.2* gives transport costs as a percentage of production value (Australian Bureau of Statistics, 1997). Wholesale trade has the highest trans-

Table 21.2. Employment locations and transport costs.

Industry group	Average distance of workers from capital (km)	Effect of distance from capital	Distance of peak from capital (km)	Percent workers	Transport costs as percentage of production
Agriculture, forestry, fishing	179	Rising		4.0	3.8
Mining	81	Peaked	171	0.3	1.9
Manufacturing	49	Peaked	30	16.8	3.2
Electricity, gas, water supply	91	Peaked	135	0.7	2.5
Construction	59	Peaked	41	6.1	2.8
Wholesale trade	52	Peaked	25	6.3	4.2
Rail trade	60	Peaked	29	14.3	1.0
Accommodation, cafes, restaurants	64	Peaked	3	4.0	1.7
Transport, storage	51	Peaked	29	4.0	4.0
Communication services	44	Peaked	12	2.3	3.1
Finance, insurance	39	Peaked	4	4.2	0.7
Property, business services	38	Falling		10.2	1.1
Government administration, defense	63	Peaked	27	3.9	3.0
Education	58	Peaked	5	7.2	0.5
Health, community services	59	Peaked	4	9.6	1.4
Cultural, recreational services	43	Falling		2.4	2.4
Personal, other services	54	Falling		3.6	2.3
Total	58			100	2.4

port costs for any industry group, but these are only 4.2% of production. For all but a few specialized industries (such as brick manufacture), transport costs may play very little role in the choice of location.

21.3.4 Location patterns for each industry

For each of the 17 industry groups, the numbers of workers per resident in each of the 195 areas were calculated, and divided by the same ratio for Victoria as a whole. The resulting values were plotted against distance from Melbourne, the capital city, and simple mathematical distributions fitted. In general, a gamma distribution plus a constant gave an intuitively acceptable fit. These industries are shown in *Table 21.2* as "peaked," together with the distance from the capital where the peak occurred. For agriculture, forestry, and fishing, employment ratios increased with distance from the capital. For three industries (property and business services, cultural and recreational services, and personal and other services), a negative exponential plus a constant gave a more plausible fit.

21.3.5 Primary industries

Agriculture is clearly a regional industry, occurring where soil and water supplies allow economic production. Some land may be unavailable for agriculture, being used for more profitable purposes, or set aside for conservation. Different types of agriculture require more labor, helping explain some of the wide variability observed between the 195 regions. Some statistical local areas are essentially rural towns, and others are the agricultural areas surrounding those towns. Some primary production appears to occur in the central business district of Melbourne, as a result of the location of head offices of agricultural firms there. In fitting models, the central business district was omitted, to avoid distortions from this "head office" effect. The fitted curve for agriculture, forestry, and fishing is a gamma function, but the wide dispersions make its validity very dubious.

21.3.6 Manufacturing

The fitted curve (a gamma function plus a constant) peaks at 30 km from the center of the city. There is high dispersion, but the highest concentration of manufacturing worker is only 187%, compared with 1,593% for agriculture, forestry, and fishing. Again, there is an anomalously high value at the central business district. The high levels of manufacture 20 to 40 km from Melbourne may reflect relatively cheap land, labor availability, and convenient access to Melbourne's consumers and transport facilities. The availability of education, entertainment, and medical facilities for workers and their families may also be relevant.

Marshall (1920) suggested reasons for the concentration of similar industries in an area:

> "Employers are apt to resort to any place where they are likely to find a good choice of workers with the special skill that they require . . . "
>
> " . . . the economic use of expensive machinery can sometime be attained in a very high degree in a district in which there is a large aggregate production of the same kind."
>
> "Good work is rightly appreciated, inventions and improvement in machinery, in processes and the general organization of the business have their merits promptly discussed . . . "

History also plays an important role. Large manufacturing plants tend to be built where land is available, at the outskirts of the capital. As the capital grows, these plants may lie well within the current sprawl, rather than at its edge.

21.3.7 Wholesale and retail industries

Wholesale trade shows a very similar pattern to manufacturing, with the fitted curve having a peak at 25 km. Wholesale trade may be concentrated close to Melbourne, to service manufacturers, builders, and the retail trade. Proximity to sea, road, and rail transport may also be important. Retail trade shows a much flatter pattern than wholesale, with a small peak at 29 km. The average distance of retail workers from the center of Melbourne is 60 km, identical with the average distance of Victorian residents from Melbourne.

21.3.8 Service industries

Property and business service workers show a very different pattern, declining sharply from the center of the city. The average distance of workers from the center is 38 km, less than any other industry group. Health and community services are almost evenly spread throughout the population of Victoria, with an average distance of 59 km from the center of Melbourne. The fitted distribution is a large constant, plus a small gamma distribution, peaking at 4 km from the center. This small peak may reflect specialized medical facilities, and the major teaching hospitals, close to Melbourne.

21.4 House and Land Prices

21.4.1 House prices

Victorian house prices are available from sale records maintained by the Valuer General Victoria (Department of Natural Resources and Environment, 1999). They are for the 78 local government areas, ranging in population from 3,500 to 184,000 persons. Mean house prices in 1998 ranged between \$47,500 and \$477,000, a range of 10 to 1. A good fit was obtained with the formula[1]

$$\ln(\text{price}) = 12.95 - 0.291 * \ln(\text{km from capital}).$$

21.4.2 Residential land prices per square km

Residential land prices, for blocks up to 2,000 square meters (m^2), were obtained from the same source as the house prices. They ranged from \$31 to \$836/m^2, a range of 27 to 1. A good fit was obtained with the formula

$$\ln(\text{price}) = 6.48 - 0.48 * \ln(\text{km from capital}).$$

[1] In the following formulas ln(x) denotes natural logarithm of x.

Effectively, land at twice the distance from the capital is worth 28% less. The reasonable fit to this simple formula, over such a wide price variation, suggests that simple mechanisms are at work. The wider range of jobs in the capital, and the better education, entertainment, and medical facilities there, all make distance from the capital a crucial determinant of residential land prices.

21.4.3 Hobby and beef farm prices

The same source provides details of sales of different types of farm land. “Hobby” farms are generally of a few hectares (ha), usually too small to allow any sort of economically viable farming. Their mean price per hectare is usually much higher than that of beef farms (this may partly reflect the inclusion of a dwelling in the price). In 1998 the total area of hobby farms sold in Victoria was recorded as 2,689 ha. Assuming the 7.2% sale rate applying to Victorian private buildings in 1998 suggests that the total area of hobby farms may have been about 37,000 ha. This is small in relation to the 12.8 million ha of agricultural land in Victoria, so that the loss of production as a result of hobby farms is still small. The value of land reduces sharply with distance from the capital, both for hobby farms and beef farms.

21.5 Migration Behavior

21.5.1 Probabilities of migration within Australia

Cumpston and Sarjeant (1998, p. 7) calculated the numbers of persons found to have changed their statistical division of residence between 1991 and 1996, as percentages of the numbers resident in 1991. Probabilities of migration were about 22% for those aged 20–29 at the start of the 5 years, and reduced to about 5% for older persons.

21.5.2 Distance-tolerance when migrating within Australia

Models of the following form were fitted separately for each age group:

ln (probability of migration from one SD to another)

= a + b1 ∗ ln(source size) + b2 ∗ ln(destination size) + c ∗ ln (km)
+ d1 ∗ ln(source unemployment) + d2 ∗ ln(destination unemployment)
+ e (latitude change) + f (longitude change) + g (1 if ocean destination, 0)
+ h1 ∗ ln(source income) + h2 ∗ ln (destination income) + i ∗ ln (source sex ratio).

At most ages, a doubling in distance roughly halves the probability of migration. Older persons are however less sensitive to distance. Although they seldom migrate, they may move long distances when they do.

A broadly similar study by Flood *et al*. (1991) noted that the regression coefficients would have changed a good deal if certain of the variables had been removed, as a result of correlations between the independent variables. The very large differences between parameter estimates from the two sources suggest that the choices of explanatory variable, and the fitting procedure, are of great importance.

21.6 Past, Present, and Future

21.6.1 The view in 1961

The models described in this chapter are similar to those proposed by Orcutt (1957) and implemented for 4,580 households by Orcutt *et al*. (1961). They commented (p. 6):

> "Our socio-economic system is a complicated structure containing millions of interacting units, such as individuals, households and firms. It is these units which actually make decisions about spending and saving, investing and producing, marrying and having children. It seems reasonable to expect that our predictions would be more successful if they were based on knowledge about these elemental decision-making units - how they behave, how they respond to changes in their circumstances, and how they interact."

Citing actuarial use of mortality data to estimate probabilities of death for individuals of given sex and age, they suggested (p. 17) that:

> "An actuarial approach also seems appropriate to predicting many types of economic behaviour . . . It may be feasible . . . to estimate probabilities of house purchase for groups of households with certain observable characteristics."

They noted (p. 399) that

> "The required computations will be enormous, but two different types of development will serve to keep computing requirements completely manageable. In the first category of developments we must list developments in computing technology. This technology is proceeding at an incredible rate and promises to continue doing so for several years . . . "

> "The second category of developments that will keep computing requirements manageable has to do with model formulation and programming . . . "

21.6.2 The view in 2000

These early ambitions have only been partly realized. Microsimulation models have been extensively used in government policy making, for example in determining taxation scales, student assistance schemes, and social security benefits (see Orcutt *et al.*, 1986). Although Eliasson (1977) modeled some actual Swedish firms, little use seems to have been made of microsimulations of firms or of regions. An Australian dynamic model currently under development (King *et al*., 1999) is based on a 1 in 100 sample of Australian census returns, subdivided by state. The Canadian Integrated Land Use, Transportation, and Environment microsimulation modeling system (ILUTE), is developing simulations of residential housing markets (Miller and Salvini 2000).

21.6.3 Computer needs

The computational requirements depend on the number of individual people and firms needed for reasonably reliable estimates. For government policy decisions affecting the whole of Australia, a 1 in 100 model is likely to be ample. For projections of the 2,600 postal areas, however, each person may have to be simulated. To keep simulation times reasonable for such models, all the data should be stored in computer memory. This will need 5 to 10 gigabytes of memory (for a parsimonious model), which should soon be available in personal computers. It is less clear that computational times will be reasonable, particularly if adequate simulations are made of the housing and employment markets.

21.6.4 Programming techniques

Klosgen (1983) described the hardware and software used in some of the major models. He commented that the limited success of MASH, the first general purpose microsimulation software system, was due mainly to orientation to a special hardware which was soon rendered ineffective by technological progress. Muller (1983) noted the desirability of using data base management systems to avoid programming restrictions and maintain data integrity. To retain flexibility and minimize programming effort, object oriented code and database systems are being used here. At least one model has suffered problems with poor program documentation, and all models and parameter estimation processes should be documented fully before programming. All systems are likely to be rewritten as experience is gained and wider data included.

21.7 Data Sources and Synthesis

Very good small-area data are available from the Australian Bureau of Statistics, derived from population censuses, birth and death notifications, business registrations, and surveys. These data are available for statistical local areas, and the Bureau provides facilities to approximately estimate postcode values. Taxation, social security, and health data are available on a postcode basis. House prices are available from state sources or commercial web sites. Additional tabulations can be bought from most of these sources.

Paass (1983) described a variety of techniques to link data samples. ILUTE is using one such method, described by Beckman *et al*. (1996), to create weighted samples of households conforming with census small-area marginal totals. Australian data samples available for research use have severely limited geographic identification, are at best 1% samples, and are costly. They are derived from census or survey data, and may poorly represent some sectors (such as the affluent). To provide the capability to model each person, and to fully use the wealth of data available, hypothetical persons and households are being created by sequential synthesis, using small-area totals and broad statistical distributions.

Synthesis starts with the estimated resident population, by sex and 5-year age groups, in a small area. Random selection is used to allocate an exact age to each person. National probabilities of partnership at each age are used to randomly determine whether each person is in a partnership, adjusting the results to give the known numbers of persons in partnerships. Males and females selected as partnered are then randomly paired off with each other, using the national distribution of age differences between partners. Similar statistical processes are used to allocate children to couples and sole parents, and to create group households. Incomes and assets are allocated sequentially, again using known totals for the small area, and national distributions.

This sequential process presents many challenges. For example, suppose that the numbers of workers resident in each small area, the numbers of persons employed in each area, and the geographic centroid of each area are known. We also have survey data on the distribution of distances between employment and residence. How are persons to be allocated to jobs? It is easy to allocate most of the employed to available jobs, using the distance distribution, but problems develop at the end of the process, when the remaining jobs may be long distances from some unallocated workers. A potential answer to such mismatching problems may be to let the projection processes simulate the movement of households to correct initial misallocations. Tests are needed on the validity of the synthesized data for particular applications.

21.8 Environmental Effects

21.8.1 Applications of employment-driven population models

The employment-driven nature of the models gives them the capacity to respond to new developments, and to project the slow outcomes of past changes such as rural mechanization. Assumptions about future industrial patterns are needed, and will inevitably be speculative. The use of microsimulations for regional projections is still new, and their reliability will be suspect for many years. Some examples of potential applications are discussed below.

21.8.2 Effects of Internet use

In theory, the Internet should reduce the need for centralized shops and offices, allow more widely dispersed populations, and reduce travel costs. In practice, these effects may be small or delayed. Only a minority of goods can be bought without visual inspection or sales assistance. The property and business services sector has grown strongly in the last decade, partly to help develop Internet and communication services. This sector tends to locate very close to capital centers, and may be partly responsible for the large growth in inner city residential prices that has occurred in most Australian capitals. The Internet does make teleworking more feasible, but social factors are likely to limit the extent to which this occurs.

21.8.3 Effects of oil price increases

Large oil price increases will reduce the willingness of workers to travel long distances to work by car. Inner city residential densities are likely to increase, with greater use of public transport. House prices in outer areas are likely to decline, and developers be less willing to build houses on green sites. Locations for new plants may be chosen with more regard for transport costs. Alternative fuel systems, with lower pollution levels, are likely to be developed commercially. The environmental effects will generally be good.

21.8.4 Effects of changes in rural industries

Australian rural industries receive little government assistance, and as a result have to adapt quickly to technological and market changes. Improvements in machinery and farming practices have helped farmers survive lower commodity prices, but have reduced labor requirements and rural populations. Major changes in the uses made of particular land types can occur quickly. For example, sheep and cattle farms in western Victoria are currently being sold for use as blue gum plantations. The labor effects are uncertain, as there do not appear to be systematic studies of the labor requirements of many rural industries. Population declines may be slower than expected, as displaced farmers may be reluctant to leave the district.

21.8.5 Environmental conflicts

Like other countries using irrigation, Australia has growing salinity problems. Solutions to these problems can be complex and costly, with long response periods. Major changes may be needed to crop types, water reticulation systems, and water prices. These changes may be needed far from the areas experiencing the most severe immediate problems. Long-term analyses of population effects may be an essential part of designing technical measures and gaining popular acceptance.

21.9 Conclusions

Although potential uses of stochastic population projections have long been recognized, little practical progress has been made. Adequate computers are becoming available, but many issues of technique remain unresolved. Some progress is currently being made in Canada and Australia, countries with good national data. The long-term nature, detailed regional focus, and economic content of stochastic projections may make them particularly helpful for environmental issues.

References

Australian Bureau of Statistics, 1998, *Population Projections 1997–2051*, catalogue no. 3222.0, 14 July, Act 2617, Belconnen, Australia.

Australian Bureau of Statistics, 1997, *Australian National Accounts – Input-Output Tables 1993–1994*, catalogue no. 5209.0, 13 June, v + 103, Act 2617, Belconnen, Australia.

Australian Bureau of Statistics, 1995, *Travel to Work, School and Shops: Victoria, October 1994*, catalogue no. 9201.2, 24 August, iv + 24, Melbourne, Australia.

Beckman, R.J., Baggerley, K.A., and McKay, M.D., 1996, Creating synthetic baseline populations, *Transportation Research A*, **30A**(6):415–429.

Cumpston, J.R., and Sarjeant, H.B., 1998, Models of Migration Between Australian States and Statistical Divisions, Paper delivered to the 1998 National Conference of the Australian Population Association.

Department of Natural Resources and Environment, 1999, *A Guide to Property Values 1998/1999*, Landata, Melbourne, Australia.

Eliasson, G., 1977, A micro simulation model of a national economy, in B. Bergmann, G. Eliasson, and G. Orcutt, eds, *Micro Simulation – Models, Methods and Applications*, Industrial Institute for Economic and Social Research, Stockholm, Sweden, pp. 179–260.

Flood, J., Maher, C., Newton, P., and Roy, J., 1991, The Determinants of Internal Migration in Australia, Report by the CSIRO Division of Building, Construction and Engineering for the Indicative Planning Council for the Housing Industry, April, Melbourne, Australia.

King, A., Baekgaard, H., and Robinson, M., 1999, DYNAMOD-2: An Overview, Technical Paper No. 19, National Centre for Social and Economic Modelling, December, vi + 48, Canberra, Australia.

Klosgen, W., 1983, Software implementation of microanalytic simulation models: State of the art and outlook, in G.H. Orcutt, J. Merz, and H. Quinke, eds, *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, Netherlands, pp. 475–491.

Lutz, W., Sanderson, W., and Scherbov, S., 1996, Probabilistic population projections based on expert opinion, in W. Lutz, ed., *The Future Population of the World – What Can We Assume Today*, revised edition, Earthscan Publications, London, UK, pp. 397–428.

Marshall, A.J., 1920, *Principles of Economics*, Macmillan, London, UK.

Miller, E.J., and Salvini, P.A., 2000, The Integrated Land Use, Transportation and Environment (ILUTE) Modelling System: Description and Current Status, Presented at the 9th International Association for Travel Behaviour Research Conference, July 2-5, Gold Coast, Queensland, Australia.

Muller, G.P., 1983, Data structure requirements of microanalytic simulation models, in G.H. Orcutt, J. Merz, and H. Quinke, eds, *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, Netherlands, pp. 495–516.

Orcutt, G.H., 1957, A new type of socio-economic system, *Review of Economics and Statistics*, **58**:773–797.

Orcutt, G.H., Greenberger, M., Korbel, J., and Rivlin, A., 1961, *Microanalysis of Socioeconomic Systems: A Simulation Study*, Harper and Brothers, New York, NY, USA, pp. vxiii + 425.

Orcutt, G.H., Merz, J., and Quinke, H., eds, 1986, *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, Netherlands, pp. xvii + 553.

Paass, G., 1983, Statistical match: Evaluation of existing procedures and improvements by using additional information, in G.H. Orcutt, J. Merz, and H. Quinke, eds, *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, Netherlands, pp. 401–422.

Chapter 22

Identification of Sustainable Forest Management Practices: Siberian Forest Case Study

Matti Flinkman, Wojtek Michałowski, Sten Nilsson, Roman Słowiński, and Szymon Wilk

Abstract

This chapter describes the identification of practices that are considered essential for the development of sustainable forest management in the Siberian forests. This identification begins with an analysis of the net primary production of phytomass, which is used to classify the Siberian ecoregions into compact and cohesive phytomass production classes in an attempt to recognize various degrees of ecosystem functioning in the ecoregions. In an earlier study, Rough Sets analysis was used to mine the Siberian forest data. This application resulted in the definition of dominant forest attributes and development of the relationships between various forest characteristics expressed in terms of so-called interesting rules. Here, we use information provided by these rules to portray forest management conditions desirable for the support of ecosystem functioning in the ecoregions. The results are presented in the form of GIS maps, illustrating the current potential of each ecoregion to enhance ecosystem functioning through various forest management practices.

Keywords: Ecosystem functioning, forestry, sustainable forest management, rough sets, GIS maps.

22.1 Introduction

This chapter deals with the identification of sustainable ecosystem management conditions for the Siberian forests, by using information from a comprehensive database maintained at the International Institute for Applied Systems Analysis (IIASA, 1998). The analysis relies on *ecosystem functioning* (EF) as a core analytical concept, implying that the appropriate and desirable operation of all facets of forest management is necessary to support ecosystem services. Delivery of ecosystem services involves: (1) capture of solar energy and conversion into biomass that is used for food, building materials and fuels; (2) breakdown of organic wastes and storage of heavy metals; (3) maintenance of gas balance in the atmosphere that supports human life: absorption and storage of carbon dioxide and release of oxygen for breathable air; and (4) regeneration of nutrients in a form essential to plant growth, e.g., nitrogen fixation and movement of those nutrients.

The analyzed forest ecosystem conditions are chosen from abiotic, biotic, and human induced factors, and thus encompass and describe the interactions between land-uses, vegetation types, forest density, site-class, age, and different aspects of human activities. The identification of conditions associated with EF is considered to be a first step towards establishing sustainable forest management practices. It is this step that is described in the chapter in greater detail, with special emphasis on those ecoregions characterized by a high level of EF.

The data component of the study is described in Section 22.2. In keeping with the idea of a comprehensive approach, it is necessary to evaluate different roles the attributes are playing in describing ecosystem conditions. The decision as to which attributes should be selected for analysis is therefore a complex one. This task was accomplished using a Rough Set (RS) methodology (Pawlak, 1991; Słowinski, 1992) enhanced by a heuristic evaluation of the possible sets of attributes with similar descriptive accuracy. The methodology was described in Flinkman *et al*. (2000). In Section 22.3, we describe how the results of this earlier study are used to evaluate forest management options. The results of this evaluation and discussion are presented in Section 22.4 and the conclusions are given in Section 22.5.

22.2 Siberian Forest Database

The Siberian forest database contains information on nearly 5,000 attributes. The spatial coverage of the collected information is aggregated at different levels. The highest level covers the whole of Siberia, while the sub-levels are for 65 ecological regions (ecoregions), 360 landscapes, and 2,500 forestry enterprises. All database items can be related to some spatial aggregation level, thus allowing for spatial descriptions of abiotic, biotic, and anthropogenic conditions.

For purposes of this paper, data aggregated at the ecoregion level were extracted from the Siberian forest database for 31 different ecosystem attributes. Some of these attributes were the result of earlier evaluations producing so called CODE-descriptors and SHDI-descriptors, describing the structure of certain forest compositions and area distributions. In creating the CODE-descriptor, the original distribution data (for example, the age distribution of a forested area) have been categorized into a few (4–7) share classes. This allows the creation of a number of distribution "profiles." The SHDI-descriptors were created based on the Shannon diversity index (Shannon and Weaver, 1962). The SHDI-descriptor represents the degree of diversity of the attribute under consideration. For example, an attribute with only a few dominating classes results in a low diversity value for the SHDI-descriptor, while an evenly distributed share results in a high value.

22.3 Problem Analysis

This chapter is based on the hypothesis that the classification of Siberian ecoregions into different classes based on the net primary production of phytomass (NPP) will reflect different types of land-use and biogeophysical conditions (Shvidenko *et al*., 1997). The net primary production of phytomass was calculated according to Bazilevich (1993) and represents a theoretical measure of an ecoregion's ideal production potential of phytomass in a particular year. This NPP measure includes all land uses, including agricultural land, within an ecoregion. Therefore, such a classification should capture a number of factors associated with the level of EF. We have assumed three levels of NPP classification of the Siberian forests. The relevant geographical information system (GIS) map is shown in *Figure 22.1*.

22.3.1 The NPP evaluation problem

Evaluation of the attributes characterizing Siberian ecoregions and a subsequent reduction of their number with the help of RS analysis was described in Flinkman *et al*. (2000). In the current chapter, we show how the results of this earlier analysis can be used to associate desirable forest structure (FS) with the NPP evaluation in order to propose forest management practices that support NPP productivity. Therefore, we focus on a reduced set of attributes, as identified by the earlier RS analysis. This set consists of 6 (out of 31 original) attributes: *relief conditions* (MOUNTAIN); *snow cover conditions* (SNOW_COVER); *share of forested area of total ecoregion area* (FA/Area); *forest fund profile (consisting of forest land,*

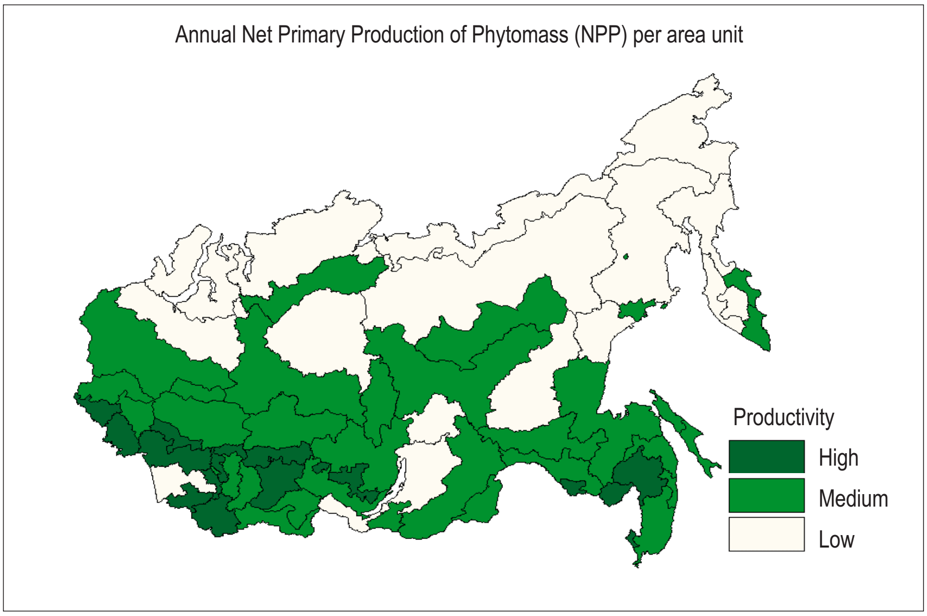


Figure 22.1. The NPP evaluation of Siberian ecoregions.

non-forest land and lease) (FF-CODE); *age profile of growing stock consisting of 5 age class categories* (AgVo-CODE); and *density of railway network* (Railw/sqkm).

The *age profile of growing stock, share of forested area of total ecoregion area*, and *forest fund* are all forest-related attributes. The *relief conditions* and *snow cover conditions* describe biogeophysical conditions, and the *density of railway network* can be considered an indicator of ecoregion development.

In the next step we generate decision rules for the reduced set of attributes using the Explore algorithm (Stefanowski and Vanderpooten, 2001). The Explore method is based on the breadth-first search strategy with the search space narrowed down by optional constraints. These constraints are imposed on generated rules and refer to their basic properties, e.g., length (number of conditions in the rule) or strength (number of examples satisfying the rule). The decisionmaker is able, unlike with other methods, to generate only those rules that satisfy his/her requirements by defining and manipulating the constraints. Thus, s/he focuses on areas of specific interest to him/her. This is why the rules generated by this method are often denoted as interesting rules. In this paper we defined a constraint specifying the minimum relative rule strength as a threshold of 10%. (Relative rule strength is defined as the ratio of number of objects satisfying the rule to the number of all objects belonging to the decision class pointed by the rule.)

Table 22.1. Interesting rules for the NPP evaluation problem.

Rule no.	NPP class	Elementary conditions AgVo CODE	FA/Area	FF-CODE	MOUNTAIN	Railw sqkm	SNOW_ COVER	Relative rule strength
1	L	AABAF						12%
2	L		0				LONG	56%
3	L			ECA	1		LONG	12%
4	L			ECA			LONG	12%
5	M	ABDBC			1			10%
6	M	AABBE			2			16%
7	M	ABDBC	1					10%
8	M		1	ECA		1		13%
9	M		1	GAA		0		13%
10	M	AACBD	1			1		10%
11	M		1	FBA	2		SHORT	10%
12	M		1		2	0	SHORT	16%
13	M			FBA	2	0	SHORT	10%
14	H		0	FBA	2			19%
15	H			FBA	2	1		19%
16	H		0	FBA		1		25%

22.3.2 Analysis of interesting rules

The *interesting rules* are presented in *Table 22.1* with each row representing one decision rule. The conditional part of the rule is a conjunction of elementary conditions on those attributes for which values are specified (the elementary condition has the syntax *attribute* = value). The decision part reflects the assignment of an ecoregion to the specified NPP class. As an example, rule 5 should be read as:

if AgVo-CODE equals ABDBC and MOUNTAIN equals 1, **then** the NPP class is M.

Values for the AgVo-CODE attribute represent different distributions of growing stock into age classes (youngest forest, young forest, middle aged forest, immature forest, and mature and overmature forest). An interpretation of this rule is as follows: **if** the distribution of growing stock into age classes is such that 0–5% of the growing stock is in the age class “youngest forest,” 5–20% is “young forest,” 40–60% is “middle aged forest,” 5–20% is “immature forest,” 20–40% is “mature and overmature forest,” and relief conditions are mountainous, **then** the NPP class is medium.

The relative rule strength for this rule (presented in the last column) is equal to 10%. This means that 10% of examples in the NPP medium class satisfies (and supports) it.

The *forest fund profile* (FF-CODE) appears to be the most frequent attribute present in the condition part of the *interesting* rules. This is particularly true for the high NPP class, where it appears in the conditional part of all decision rules. For the two other NPP classes, it appears in combination with a number of other attributes in most of the rules.

Values of the FF-CODE attribute represent different distributions of land within forest fund into land use classes [forestland, non-forested lands, and "long-term lease lands." The letter gives the share percentage ranging from $<$5% (A) to $>$95% (G)].

In addition, values of the SNOW_COVER attribute reflect different duration of a snow cover, with value LONG denoting long winter and SHORT denoting short winter. Values 0 and 1 for the attributes FA/Area and Railw/sqkm indicate either the first or second interval generated by the *Recursive Minimal Entropy Partitioning* discretization method (Fayyad and Irani, 1993) applied for these two attributes. All other attributes were discretized according to the value intervals provided by an expert.

22.4 Results and Discussion

In light of the complementary character of the attributes *forest fund profile* (FF-CODE) and a *share of the forested area of the total ecoregion area* (FA/Area), we conducted a further evaluation of the interesting rules associated with the high ("H") NPP class only. This evaluation necessitated different interpretations of the rules, and resulted in the establishment of the so-called *FF-interpretation* and *FA-interpretation* of the *interesting* rules, as defined below.

22.4.1 FF interpretation

The *FF-interpretation* relies primarily on a *forest fund profile* attribute while describing good EF.

In general, the ecoregions with a forestland being a dominant part of the forest fund, characterized by developed infrastructure and plain relief conditions, have good EF. Moreover, a typical feature of the ecoregions identified here is that the forests cover only a minor part of the ecoregions' total area. This implies that good EF depends mainly on life forms other than forests. A common characteristic of the forests in these ecoregions is a relatively large share of growing stock in the fast growing, middle age classes, resulting in high productivity in the forests.

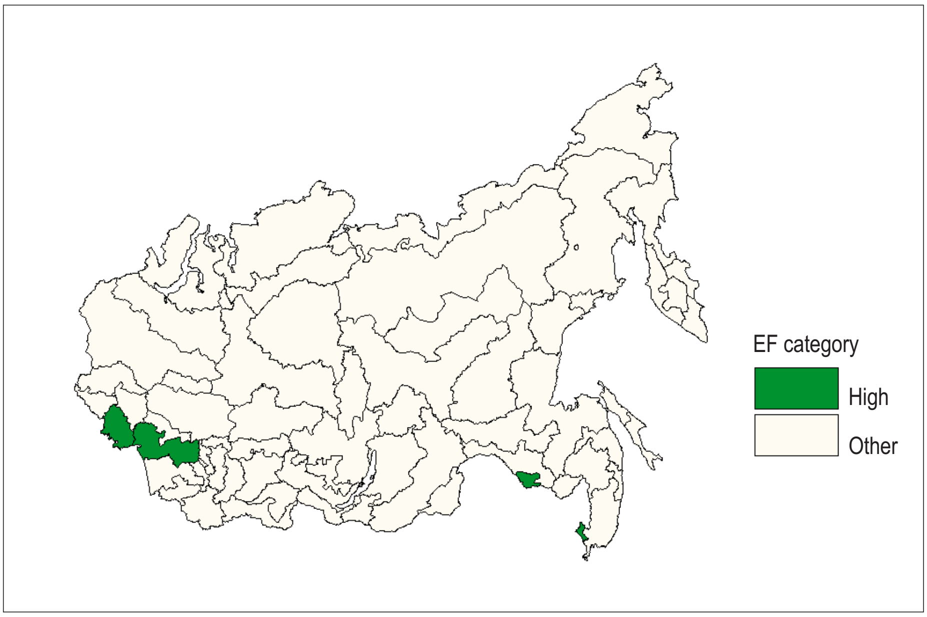


Figure 22.2. The FF-interpretation.

Clearly, the application of the *FF-interpretation* identifies those ecoregions that show the absolutely highest NPP among all Siberian ecoregions. It is illustrated in *Figure 22.2*.

22.4.2 FA interpretation

The *FA-interpretation* relies primarily on a *share of the forested area of the total ecoregion area* while describing good EF.

Ecoregions belonging to this category are presented in *Figure 22.3*. Clearly more ecoregions are identified here, and the conclusion is that ecoregions characterized by a low share of forested area, a developed infrastructure, and plain relief conditions have good EF. In addition, a typical feature of these ecoregions is that the share of forested area of the total ecoregion area is, on average, higher than those selected with help of the *FF-interpretation*.

As exemplified by the above analysis, the extent and condition of forests may play a crucial role in determining the level of NPP and EF. The role of the forests in the context of EF is, however, not straightforward. The relatively high share of forested areas can be advantageous or disadvantageous in promoting EF, depending on the type, character and condition of the forest stands. Specifically, those ecoregions with a less favorable age structure, that were identified as having good EF according to the *FA-interpretation* only (they are not identified as such when the

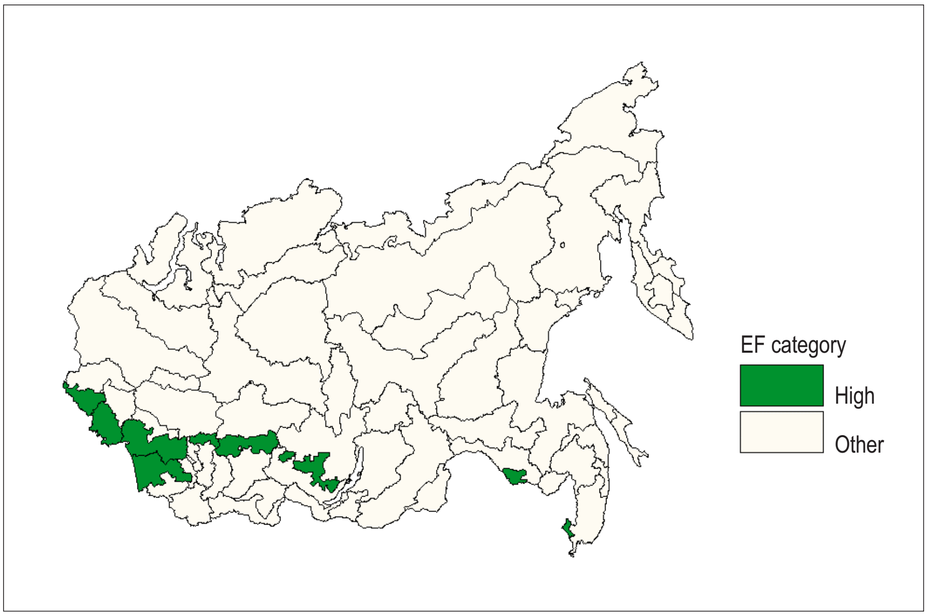


Figure 22.3. The FA-interpretation.

other interpretation is applied), represent the potential to increase NPP and improve EF through the implementation of more efficient forest management practices (for example by cutting and replacing the large share of "old growth" forest with fast-growing young stands).

In summary, the ability to associate a specific set of dominant characteristics (e.g., the distribution of growing stock into various age classes) with the ecoregions characterized by "high" NPP and EF, allows us to establish and recognize a *potential* of the forests to enhance EF by implementing alternative forest management practices. In this context, it is of interest to distinguish ecoregions with *desirable forest structures* (FS) originating from specific types of forest management practices, and/or natural disturbances, from those having *undesirable FS*, due to less sustainable management practices and/or lack of natural disturbances in the past. In the following example, the analysis of a reduced set of 6 attributes, and associated *interesting* rules, is applied to discover directions for alternative forest management practices. Such analysis suggests that the *desirable* FS be:

- characterized by relatively low snow coverage of the landmass,
- associated with the forest fund having more than 60% of forestland, and
- characterized by forests with a lower share of growing stock in mature and overmature age class, as well as in the immature age class.

In general, this characterization of desirable FS patterns will identify ecoregions having favorable growing conditions, shallow snow cover, and productive forests. All other ecoregions not exhibiting these characteristics would have the undesirable FS pattern. Such a broader interpretation of the EF also means that the ecoregions belonging to any of the NPP classes may exhibit desirable FS.

Therefore, while considering the FS, one evaluates ecoregions based on the degree of forest management impact on the overall EF. In the ecoregions with the desirable FS, the potential to improve EF through forest management practices is lower than for those with the undesirable FS. In the latter, implementation of different forest management actions might improve their FS, for example, through a reduction of the share of " old growth" forests. So the desirable or undesirable FS patterns not only express the degree of EF, but also capture the potential of forest management actions for enhancing the productivity of the forests.

Having distinguished desirable and undesirable FS, we imply that ecoregions with desirable FS are controlled, to some extent, by man or natural disturbances that ensure a relatively low share of growing stock in immature and mature and overmature age classes, and subsequently exhibit less potential to increase productivity (NPP) and EF. This thereby allows us to examine the forest management practices associated with the ecoregions of desirable FS, and to transplant them to those ecoregions that have undesirable FS patterns. The underlying assumption is that ecoregions with undesirable FS patterns are not well managed by humans, and/or there have been few natural disturbances in the past, resulting in a high share of growing stock in " immature" and " mature and overmature" age classes. Thus, they should present a higher potential to increase productivity and EF.

The ecoregions with desirable FS, and thus a lower potential to improve their productivity (NPP), are presented in *Figure 22.4*. When it is compared with *Figure 22.1*, it is clear that some ecoregions with medium or low NPP are now in fact identified as having desirable FS. The ecoregions with desirable FS have accordingly a lower potential to increase their productivity (NPP), even if more effective forest management was applied. This is specifically the case with these ecoregions not belonging to the " high" NPP class. On the whole, the ecoregions with desirable FS are considered to be relatively well managed in terms of utilizing the forest productivity potential.

22.5 Conclusions

In this chapter we demonstrated how to identify management practices for the development of sustainable forest management policies in the Siberian forests. Our analysis has resulted in the identification of the dominant ecosystem characteristics

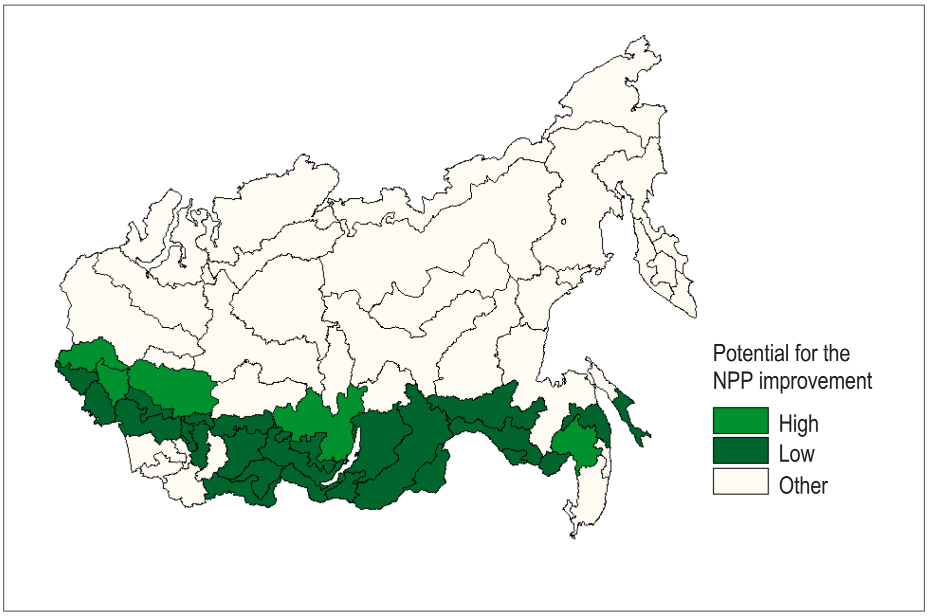


Figure 22.4. Potential for the NPP improvement.

and the assessment of relationships between these characteristics and EF. This allowed articulation of some forest management practices for maintaining desirable FS in order to enhance sustainable delivery of ecosystem services. We thus analyzed the interactions between dynamic EF and biotic, abiotic, and human-induced conditions by:

- assessing common EF circumstances of Siberian ecoregions and illustrating them in a form of GIS maps (following a framework given in Flinkman [1999]),
- identifying a dominant set of ecosystem conditions, and
- identifying forest management practices that may enhance a sustainable and desirable EF from the point of view of anthropogenic aspects.

Furthermore, the results allowed us to propose alternative forest management practices, including re-direction and improvement of current practices with respect to sustainable and desirable EF.

Our approach provided a convenient link, in this case from the forestry perspective, between the ecosystem conditions and the EF and its improvement potential through identification of those forest structures and associated forest management activities that might enhance EF. Thus, various forest management practices were

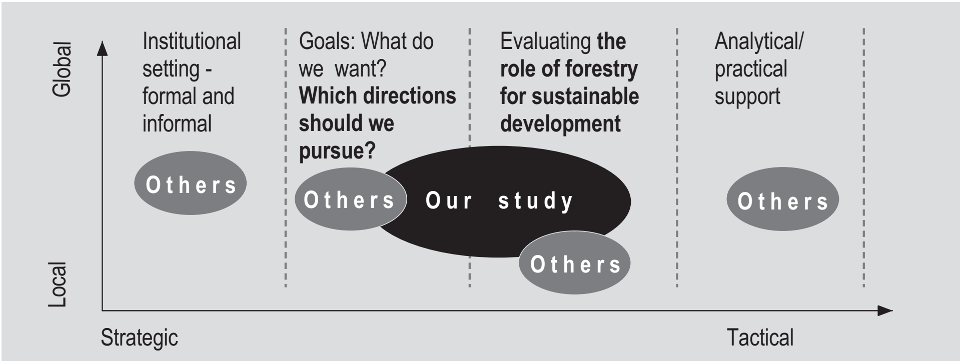


Figure 22.5. The general framework for evaluating the role of forestry in sustainable development.

scrutinized from the point of view of their potential impact on EF. The position of this paper along the goals and settings dimensions is given in *Figure 22.5*. It shows that our paper combines an evaluative component with a goal setting mechanism by associating desirable EF with focused management practices.

The proposed approach clearly has a potential to evaluate other areas where human induced impact on EF is crucial for the natural processes and delivery of ecosystem services. Simultaneously, it should be further validated by using disaggregate data at a regional/local level, including alternative forest evaluation perspectives. Results from such an analysis might be then used as a basis for recommendations relevant for practitioners operating in the field.

Acknowledgment

The authors wish to thank Ms. Anne Burgess for editorial assistance.

R. Słowinski and Sz. Wilk wish to acknowledge financial support from the Polish Committee for Scientific Research (KBN). The Decision Analysis and Support Project at IIASA and the Natural Sciences and Engineering Research Council of Canada (NSERC) supported research conducted by W. Michalowski while he was a senior research scholar at IIASA.

References

Bazilevich, N.I., 1993, *Biological Productivity of Ecosystems of Northern Eurasia*, Nauka, Moscow, Russia (in Russian).

Fayyad, U.M., and Irani, K.B., 1993, Multi-interval discretization of continuous-valued attributes for classification learning, in R. Bajcsy, ed., *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, San Mateo, CA, USA, pp. 1022–1027.

Flinkman, M., 1999, *Pan-Siberian Forest Biodiversity: Identifying Sustainable Forest Management Practices*, IR-99-030, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Flinkman, M., Michalowski, W., Nilsson, S., Słowinski, R., Susmaga, R., and Wilk, Sz., 2000, Use of rough sets analysis to classify Siberian forest ecosystems according to net primary production of phytomass, *INFOR*, **38**(3):145–161.

IIASA (International Institute for Applied Systems Analysis), 1998, IIASA Forest Database, see description at http://www.iiasa.ac.at/Research/FOR/dbdoc/index.html

Pawlak, Z., 1991, *Rough Sets. Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Shannon, C.E., and Weaver, W., 1962, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, USA.

Shvidenko, A., Nilsson, S., and Roshov, V., 1997, Possibilities for increased carbon sequestration through the implementation of rational forest management in Russia, *Water, Air and Soil Pollution*, **94**:137–162.

Słowinski, R., ed., 1992, *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Stefanowski, J., and Vanderpooten, D., 2001, Induction of decision rules in classification and discovery-oriented perspectives, *International Journal of Intelligent Systems*, **16**(1):13–28.

Chapter 23

A Spatially Explicit Market Model for Forest Sector Analysis

Michael Obersteiner

Abstract

The modeling approach taken in this chapter was originally designed to model relocation and restructuring patterns of the Siberian forest sector during its transition from a command economy to an economy guided by market principles. Given the possibility of linking to an extensive GIS-database, the modeling strategy was to build an easy-to-understand and easy-to-compute economic model that makes maximum use of existing data and expert knowledge. The core model is based on auction theory. The static phase describes the auction mechanism, which involves a number of strategy sets of unequally powerful individual agents. These are guided by multiple preferences and in addition are allowed to form network structures. In the dynamic phase, product prices adjust based on slack supply expectations, producers invest or depreciate capital, contract partners revise contract policies and, finally, factor prices adjustments are exogenous reflecting the overall economic development. Demand equations are estimated using Baysian econometrics and panel data.

Keywords: Economic geography, auction theory, simulation, quasi-optimization, timber product markets, Russia, geographic information systems, vertical and horizontal integration, inefficiency and wrong beliefs.

23.1 Introduction

Markets are central institutions of Western societies and recently they have also been the dominating institutions in post-communist societies. However, we do not know much about how they function or about their emergence. Although markets are so central to how our real economies and societies function, economists make little use of the few insights into the real patterns and determinants of markets when they model economic phenomena. How are prices formed? Which ways do buyers and sellers meet and do business? How are the determinants of market function evolving? How embedded are markets in formal and informal institutions? All these questions make an analysis of real market phenomena very difficult and complex. However, it is indispensable for an analyst to take these phenomena into account if s/he is to give an educated judgment of the status and future pattern of a particular market.

Spatial models in economics can be split into two categories: (1) econometric models explaining re-location patterns in industry (e.g., Hanson, 1998; Obersteiner, 2000) or (2) labor movement and models on understanding the optimal structure of cities (Gaspar and Glaeser, 1996). However, virtually no work has been conducted on modeling industries in geographically explicit terms, by modeling the interactions of production firms and consumers.

The main reason for choosing the auction approach was to allow modeling deviations from the optimal outcome. Without going into all the technical details, the following – sometimes crucially wrong – assumptions of equilibrium models need to be considered:

- assumptions on perfect competition and perfect information,
- identical agents (beliefs and strategies) and identical technology,
- frictionless markets and no externalities,
- no economies of scale, and
- no policy distortions.

The purpose of this chapter is certainly not to answer all the above-mentioned and many other questions on how the different markets of the Siberian forest industry function, but, more modestly, to try to incorporate some aspects of market functioning in a new economic model of production and trade. The general theoretical model was specially tailored and applied to model the Siberian forest industry during its transition to a 'market' system. Obersteiner (1998) provides a detailed description of both the model applied to the Siberian forest industry and the computed scenarios in graphical format.

Since the collapse of the socialist system in 1991, Russia has committed itself to pro-market reforms. The country is still undergoing a transition, which is characterized by the elimination of state subsidies, decentralization of decision making, privatization, and the abolition of price and wage controls. The former tradition of soft budgets, which guaranteed employment and low inflation, has been abandoned. National and international advisers have clearly failed to acknowledge the path-dependency of efficient policy measures, and frequently have confused the goals with the means.

So far, transition at the enterprise level has meant radical changes in the ownership structure, and fewer changes in the elimination of non-competitive products. This situation, combined with the collapse of COMECON and a lack of institutions and knowledge, led to perverse incentives that explain value chains of negative value added. The entire economy and the forest sector in particular have faced an unprecedented economic decline that no country in modern history has had to go through. The decline started in 1989–1990 and still continues, with the exception of some export-oriented enterprises or product lines. Productive assets are largely obsolete as already described by Voevoda (1985).

The official rate of unemployment is rather low, but in reality it is much higher. Hidden unemployment increasingly becomes apparent. Workers over the age of forty are most affected by the transition. Pensioners barely survive due to empty pension funds. The economic downturn has affected remote areas the most.

Decline in real incomes and construction activities adversely affects the consumption of forest products. Consumption around the turn of the century will most likely be 50% of what it was ten years ago. Adoption of a new economic culture securing the efficient functioning of markets will probably take generations. Well-established Mafia networks and unrealistic profit- and rent-seeking by most managers exemplify the concerns of slow adaptation. The economic problems in Russia are also connected to high political risks, which in the worst case may erupt in civil unrest due to unequal distribution of wealth or nationality conflicts. Such unforeseeable events would of course destroy all scenario exercises.

The level and geographic distribution of forest sector output will mainly depend on two factors. The first is the competitiveness of domestic woodworking industries and the second is the trade pattern (Backman, 1996). After the deep economic slump, there are few enterprises that are still competitive on domestic or international markets. Those who inherited decent working capital from the Soviet system are on the way to recovery. However, many producers had to close down operations and many more will do so in the future.

Trade links with former Soviet satellite countries such as Bulgaria, Hungary, East Germany, and the Baltics have dried up. Future delivery, especially of roundwood from remote Siberian areas to European Russia, will heavily depend on the future of transportation costs. Most of the trade is organized through barter contracts, which fail intentionally to convey economic information. However, there are promising possibilities for trading timber and timber products with some former Soviet republics and countries with wood deficits in East and Southeastern Asia and the Middle East.

Future competitiveness of the Siberian forest sector now centers around the question of how the allocation and structure of production will change during the transition to a market economy. It is not only the change in prices that will drive this development. Today, economic agents in Siberia follow economic paradigms that are different from those used under the previous regime. Different decision rules guide economic activities, due to a framework of changed constraints. Russia has also opened its borders, which allows more freedom for trade and capital flows. The interaction of these factors with many others will define the future path of the Siberian forest sector's battle for higher market shares, both nationally and internationally.

The purpose of this chapter is to propose a new analytical tool that allows us to explore the future development of the Siberian forest sector given its geography and specifics in the transition to a market economy. The auction approach was selected because it allows 'transparent' deviation from the theoretical optimal outcome. Optimal resource allocation is unlikely to happen in an immature young market economy like that in Russia. Thus, the parameterization of the auction model allows us to learn more about the factors that could explain the deviation from the optimal state predicted by theory. The identification of inefficiency-explaining variables and their qualitative description is the most important value added of the model approach presented in this chapter. Parameterization in itself is a rather difficult task, since there are very few reliable data (such as prices or production figures) that could serve as reference points to calibrate the model in such a way that it mimics the actual facts.

23.2 Descriptive Model Formulation

Let us first start with an intuitive description of the model. The first economic activity of business partners in the forest industry in Siberia is to establish trade contracts. Imagine now, for simplicity, a fictitious market hall at a trading point in Russia. Or, the image could also be a traveling sales or purchase agent of a particular organization. The trading point could be at one of the Siberian borders for export trade, or located at the border to a woodworking mill, or a final consumer of finished products within Russia. All producers who are competitive, have available capacities, and are potentially willing to sell their products are invited to this market hall. Buyers and sellers meet here. Some of them already know each other from joint business activities or have even established long-term contracts. Buyers learn about price trends and the quality of the products that the producers are trying to sell. As a side effect of this discussion, both buyers and sellers can grasp the reliability of their possible contract partners. In addition, buyers learn about the geographic location of producers, the technology used, the productive capacities installed, and other producer-related information. Armed with this information, buyers can infer the total cost of delivery of each producer, the possible grades delivered, and important preference information.

In the model specification, which was used and adapted to model the Siberian lumber market, buyers believe *inter alia*, to some degree, that:

- Reliability and quality of services increase with the profit margin each producer can potentially earn.
- Buying from a high profit producer today will guarantee cheaper contracts in the future when competition might become stiffer and prices adjust. However, relying on just a few, but permanent and low-cost, contract partners is worse than contracting with more, but slightly more cost-intensive and competing ones.
- Buyers have the tendency to contract with large (small) firms to minimize transaction costs and risk.
- At a generally low price level, producers are more likely to be uncertain contract partners, whereas the reliability of the buyers increases.
- Only producers who can profitably deliver their products can be reliable contract partners.

After all of the information has been exchanged, the auction can start. According to their preferences, buyers potentially decide where to buy and give out their orders at the base price. Due to the fact that not all producers are willing to sell off their goods at the base price, only a portion of the total demanded volume is actually contracted out at the base price. The bargaining power of buyers against sellers is modeled by setting the quantities that are given out for tender in each auction. After the buyers have revealed their preferences to the individual producer via the offered contracts, producers decide to accept, modify, or reject the proposed offers from all buyers. Producers, of course, will screen all proposals and will select among them. The most desirable (profitable) proposals are accepted immediately, whereas the less desirable will be partially accepted with the hope to return to an even better offer when tender prices have increased, but this is taking the risk of being out-competed by new entrants. Some relatively unprofitable proposals will never be accepted.

In the Russian business environment, ordered products might get lost or, to some extent, be of bad quality. Thus, less than the entire contract volume as originally negotiated is actually realized. Strategically, producers have the tendency to deliver bad quality if prices are low. Buyers, on the other hand, will postpone payment or complain about the delivered quality of the product with increasing prices.

It follows that, after the first round of negotiation, buyers with unsatisfied demand and sellers with free capacities have to meet and the same auction procedure will start all over again; this time with increased product prices and different tender volumes. It should be noted that increasing prices cause new competitors to enter the market, thereby leading to enhanced competition. All of the agents involved now analyze past market processes; update their information; and develop new strategies for investments, price, and contract policies for the next period to come. The policy setting of each individual agent involves analysis of the past as well as expected future market developments, including prices, competition, total demand, overall investment climate, and uncertainty. Strategies, if feasible in light of the financial and natural resource conditions, are implemented by each producer. Trading according to the tender mechanism starts again in the following period. If there are still volumes to be tendered in subsequent auctions, product prices rise continuously until either producers run out of capacity or resources, or buyers have satisfied their total demand.

23.3 Formal Formulation

23.3.1 Static phase

The Buyers' Problem

$$Y_{ijtk}^{B} = f(X_{ijtk}\nu_{ijkt})D_{jt}a_{jtk} \tag{23.1}$$

where:

Y_{ijtk}^{B} ... Contract potentially offered to producer i by buyer from market j at time t in k^{th} tender.

X_{ijkt} ... Matrix of decision variables for buyers of market j purchasing from producer.

ν_{ijkt} ... Preference weight for decision variables for buyers of market j purchasing from producer i in k^{th} tender at time t.

D_{jt} ... Total tender volume at market j at time t.

a_{jtk} ... Parameter determining the size of tender volumes for the k^{th} auction at market j at time t. $a_{jtk}\epsilon]0,1]$.

Function (23.1) is about the computation of Y^B, which stands for a contract proposal in physical terms by a buyer at a particular stage in the tender process. This purchase proposal is then still subjected to a number of constraints, which are briefly discussed below. The unconstrained proposal depends on the quantity-price schedule of the buyers, which is reflected by the parameter a, and the contract affinity term between buyer and seller which is defined by the the f-function. In the current version of the model the decision variables in X_{ijkt} are made up of measures like a technical cost approximation given certain technology options, the size of the firm, an indicator matrix of membership in a production network (domestic or international), current price level, or cost and contract uncertainty on both the buyer's and seller's side. Through transformation and premultiplication with the variance-covariance matrix, the decision variables are made orthogonal to each other, which allows the calculation of Euclidean distance measures. Each individual decision variable is assigned a 'shadow' value or parameter ν_{ijkt}, which allows the calculation of a contract affinity score of each buyer to each seller. The parameters can either be estimated from real data or be set by the analyst after some sensitivity analysis. The contract affinity score is computed such that $\sum f(X_{ijkt}, \nu_{ijkt}) = 1$ and is multiplied by a fraction, $D_{jt}a_{jtk}$, of the total demand in the k^{th} tender. The parameter and the price change within the simulated auction procedure $\Delta\bar{P}_{jtk} = \bar{P}_{jtk} - \bar{P}_{jtk-1}$ are used to mimic market power of the market participants. D_{jt} is predicted using Bayesian econometrics (see, e.g., Hamilton, 1994) and panel data (see, e.g., Baltagi, 1995; Maddala, 1993) from the respective regions and macro-regions.

Function (23.1) computes the unconstrained purchasing proposal (Y^B) of buyers. However, this proposal must meet certain criteria in order to be feasible. If all constraints are met, then the potential proposal Y^B becomes the feasible contract proposal Y^{B^*}. The constraints make sure that (1) there is a minimum contract size proposed by the buyers, (2) resource and production constraints are respected, (3) supply would not exceed demand, and (4) costs do not exceed the current tender price proposal. The latter can be loosened up by allowing for certain dumping behavior over a restricted period of time.

$$\begin{aligned} Y^{B^*}_{ijtk} &= Y^B_{ijtk} \quad \text{if} \quad \begin{cases} Y^B_{ijtk} & \geq \quad Y^{min}_{ij} \\ \sum_k \sum_j Y_{ijtk} & \leq \quad \bar{Y}_{it} \\ \sum_k \sum_i Y_{ijtk} & \leq \quad D_{jt} \\ C_{ijt} & \leq \quad \bar{P}_{jtk} \\ \sum_k \sum_i \sum_j Y_{ijetk} & \leq \quad \bar{R}_{et} \end{cases} \\ Y^{B^*}_{ijtk} &= 0 \quad \text{if} \quad else. \end{aligned} \tag{23.2}$$

$\bar{Y}_{it}$ ... Production capacity of producer i at time t.
Y_{ij}^{min} ... Minimum contract size. $Y_{ij}^{min} \epsilon \Re^{+}$.
C_{ijt} ... Unit production cost (CIF) of producer i at time t to market j.
$\bar{P}_{jtk}$ ... Vector of maximum total cost at market j where transaction is allowed at time t in k^{th} iteration.
R_{et} ... Biological resource constraint in ecoregion e at time t.

The Sellers' Problem

After the buyers have revealed their offers Y_{ijtk}^{B} in the k^{th} tender to their potential set of producers, each seller will individually decide which offer s/he will accept, modify, or reject. In the version applied to the Siberian forest sector, the seller will fully accept the offer if it comes from the buyer who guarantees sufficiently high returns. S/he will modify, i.e., decrease the contract size, if the contract would yield a profit below the benchmark profit and will reject it if the offer would yield profits that would not justify delivery due to the insufficient size of the contract. In the last case, the seller believes that s/he would be able to sell again at a higher yield in a tender to come. This behavior is formalized by the following:

$$Y_{ijkt} = \begin{cases} Y_{ijkt} & \text{if } \pi'_{ijkt} = \pi'^{*}_{ijkt} \\ Y_{ijkt}(\frac{\pi'_{ijkt}}{\pi'^{*}_{ijkt}})r_l & \text{if } \pi'_{ijkt} \leq \pi'^{*}_{ijkt} \text{ and } Y_{ijtk}(\frac{\pi'_{ijkt}}{\pi'^{*}_{ijkt}})r_l \geq Y_{ij}^{min} \\ 0 & \text{if } \pi'_{ijkt} < \pi'^{*}_{ijkt} \text{ and } Y_{ijtk}(\frac{\pi'_{ijkt}}{\pi'^{*}_{ijkt}})r_l \leq Y_{ij}^{min} \end{cases}$$

The prime on π'_{ijkt} indicates that the seller selects offers from feasible contract partners only, as defined by the constraint set of the buyers' decision. π'^{*}_{ijkt} indicates the benchmark profits. r is a logistic scaling factor for the modification of the revised contract size proposed by the seller. This simple one-dimensional model of quasi-profit maximization on the sellers' side can easily be substituted by any more complex goal function.

Optimal Behavior

From a theoretical point of view, the numerical calculation would converge to an optimal solution of the allocation pattern of production for the buyer if (1) the buying/selling uncertainties would converge to zero, (2) the parameter of profit maximization in the decision variable set of the buyers would approach infinity (or a more complex goal function be optimized), (3) all of the volume would be given out for tender in the first auction, (4) the price increase from one auction to the other would be infinitely small but still positive, and (5) the number of auctions within a period would be allowed to be infinitely large so that all potential capacity could be used.

Producers are allowed to optimize only within an individual tender. They are not able to optimize dynamically over all tenders within a period and are thus also not able to optimize over periods. Currently, there is work underway to prove that no analytical optimum can be computed in this setting.

Given the present uncertain economic environment and the rather poor enterprise information, the parameter values of the model will not be close to these optimal values, but will converge slowly towards an optimum. This type of model was constructed to allow agents to act according to an optimizing behavior, but only to a certain extent. Thus, in using this model, the user can work with different degrees of optimizing behavior of the agents. In the model applied to the Siberian forest sector, calibration was done by educated judgments relying on personal experience or expert consultation. Another method of calibration would be through experiments with real economic agents or by using econometric techniques.

The possibility for loosening the optimization 'constraint' is probably the most distinguishing feature of this approach compared to the standard models used in economics and operations research. This is the beauty but also the weakness of the modeling approach taken. The quality and the usefulness of the results hinge on the calibration of the model. We are truly opening Pandora's box, allowing for a lot of arbitrary inputs featured by bounded and fuzzy rationality involving multidimensional pay-off functions. The trick is to open Pandora's box in a controlled manner, which then should justify the deviation from the standard model.

23.3.2 The dynamic phase

In the dynamic phase, changes in capacity are calculated following different investment depreciation regimes, prices are allowed to adapt according to expected demand and supply scenarios, maximum potential demand is forecasted, and costs change according to input price and technology scenarios. Not all of the variables mentioned are fully endogenized. The art of making the static model dynamic is determined by transition conditions for the endogenized variables and finding the appropriate scenarios for exogenous variables.

Capacity Changes

Investment is related to (1) the current allocation of capacity, (2) the current supply/demand pattern, (3) the accumulated profits of individual enterprises or production networks, (4) current and expected returns on investment given scenarios of the long-term future industry structure, (5) strategic capital targeted to particular markets, and (6) a number of behavioral patterns like the willingness of producers to reinvest in the same activities. Technological upgrading and capacity expansion are also indirectly dependent on the economic or technical depreciation of installed capacities.

In the current version of the model, three separate investment cycles have to be considered. There are two independent investment cycles and divestment. Capacity that will be available to each producer in the upcoming period, $\bar{Y}_{it+1}$, is made up of the sum of already existing capacities, $\bar{Y}_{it}$, plus capital additions determined in the dynamic phase, $\Delta\bar{Y}^i_{it}$, minus capital assets $\Delta\bar{Y}^d_{it}$ that are depreciating due to overcapacity of a particular producer. Capacity additions or investment are again the sum of two parts. The first is capital generated by the enterprise's own profits. This is calculated by the discounted total accumulated profits targeted for a specific production technology. Profits can be reinvested in the same activity (e.g., lumber production), or in another activity within the forest industry, or may exit the forest industry. The second investment component is foreign (extra-enterprise or extra network) venture capital. The total sum of venture capital is distributed among enterprises with market potential. Market potential is measured by a comparison of the market position at time T, at a given market, and its current position. Technology is allowed to vary according to market, e.g., high tech investment for expected deliveries to Japan and low tech for deliveries to China.

$\Delta\bar{Y}^d_{it}$ defines the depreciated capital stock due to excess capacity and/or 'natural' depreciation. Excess capacity occurs if capacity that was inherited or built up over time is not employed due to the economic supply and demand scheme. Physical excess capacity depreciates at a different rate than does excess capacity and depreciation rates are different for the various technologies. In addition, capacities are bounded from below, which is motivated by the belief that an enterprise needs to be of a certain minimum size in order to stay in business.

The 'Quasi-Optimal' Supply Pattern at Time T

The calculation of the quasi-optimal state follows the same procedure as discussed in Section 23.3.1 describing the static phase. Parameters are adjusted so that the quasi-optimal state, at time T, is set according to the considerations discussed in the subsection called Optimal Behavior. The quasi-optimal state is perpetually updated. This causes some inertia in the evolution pattern of capacity building and is motivated by the fact that producers can base their decisions on information that is available at time T. Thus the effects of technological/price shocks and changes in the competitive position of enterprises can be anticipated by the agents only to a limited extent.

Overall, the set-up of the dynamics of the model is very similar to a standard inter-temporal stochastic optimization procedure solving time-continuous stochastic optimization problems. The distinguishing feature is that we do not necessarily arrive at an optimal solution in T, but there is a persisting tendency to converge to this state. Perfect matching with the optimal state is not obtained for a variety of reasons: *inter alia* natural resources or investment capital are insufficient, demand is lacking, or contract building behaves in a stochastic manner. In addition, it is difficult to converge to the optimal state, since the model set-up makes it a moving target (within certain boundaries).

Price Determination

It is generally accepted among economists that, under a certain market structure, prices need to adapt according to the interaction of current and future demand and supply. In fact, most mainstream economists believe that price rigidities explain departures from trend growth and equilibrium states. In the Russian case, prices are rather rigid due to price control and information lags. Export prices are regularly fixed by semi-governmental organizations in Russia. The model tries to account for some of these imperfections and allows for explorations of altering policy strategies specially targeted at the sector or more general economy-wide policies.

In the model, changes in prices are endogenous. The price changes are directly determined by the model variables, given some assumptions on the process of price determination. This is not a trivial task from a computational and a model-building point of view, since the price change depends essentially on all the other variables and coefficients of the model. Hence, it becomes difficult to make a correct and fast diagnosis of price fluctuations predicted by the model. In the model, price changes are modeled as a function of the expected supply slack, the expected total investments, and the current profitability level of the industry. The expected supply slack is the difference between the current supply, which is the installed capacity, and the expected demand at time t of the product concerned at market j in period $t+t^{'}$. The price changes are computed individually for each market. The logic here is that prices change according to the ratio of total expected capital requirements to total expected revenue calculated at time t in order to match the expected demand. Thus, prices adapt in such a way that there is a tendency to market clearing. If total predicted capital requirements for the upcoming period, with prediction period of $t^{'}$ years, is larger than today's total revenue, then prices increase, and vice versa.

The ratio of total expected capital requirements to total expected revenue is used to calculate the change in the initial tender price in the next period $\bar{P}_{jt+1,k=1}$ and the change in the price change within the ascending price tender auction $\Delta\bar{P}_{jtk}$. The ratio is adjusted by a logistic filter. The increase in prices in the ascending price tender auction thus depends crucially on the expected net supply slack. If the expected supply slack is small, the initial price falls, and furthermore, the price change within the tender decreases. The latter affects both the variance and the mean of the prices achieved at the particular market in the particular year, which reflects the degree of competition. This assumption, however, still lacks empirical testing. Implicitly, this modeling approach allows investment and consequently prices to be determined by the strategies of the agents being active on the concerned markets. This is to say that, e.g., raising investments in Siberia targeted for the Japanese market will lead to lower and less volatile export prices for Japanese traders in the subsequent periods.

Uncertainty

The model includes uncertainty on a number of processes active in the Siberian forest sector. These are uncertainties in the production costs, price setting, and on contract reliability of Russian producers and traders. This issue is not further described in this chapter; see Obersteiner (1998).

23.4 Scenarios

Most of the emphasis so far has been on model building. The model will now be illustrated by one scenario for the sawmilling industry. First, the path of output (*Figure 23.1*) and the path of capacities *(Figure 23.2*) of individual firms over time are presented. In addition, the price path (*Figure 23.3*) for export markets is shown. Prices on the domestic market are not illustrated here. The investment path separated by foreign and domestic investment sources is shown in *Figure 23.4*. The cross-sectional results with respect to the installed capacities five periods (years) after the initial state are plotted on the geographic map of Siberia in *Figure 23.5*.

Figure 23.1 shows the supply pattern of individual firms over ten periods. The first year can be regarded as a self-calibration phase in the modeling process. This can be seen by looking at changes in the supply pattern from period one to period two. In period two there are ‘suddenly’ new enterprises popping up despite the fact that there are no investments observed (*Figure 23.4*). These new enterprises are in fact old enterprises of gigantic size (*Figure 23.2*) and can only enter later when they have resized and thus become cost competitive. This already happens in period two due to the high depreciation rate of unused capital. In the simulation it was assumed that the effective depreciation rate for these enterprises is $\delta_t \times \delta_t^* = 0.54$. This high depreciation rate is backed by the empirical fact that these giants had to resize considerably by closing production lines in the real world.

Following this somewhat artificial process in period one, we can observe two processes that drive industry evolution. The first process can be called *resizing* (*Figure 23.1*) and the second *relocation* (*Figure 23.5*). The latter describes changes in the location of production, whereas resizing can be described as adaption of existing firms to an ‘optimal’ size guaranteeing competitiveness. One can clearly

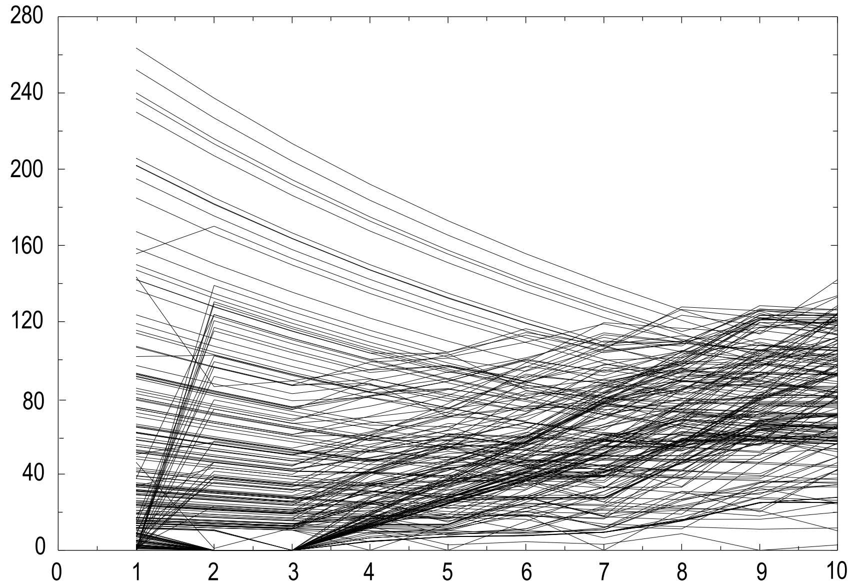


Figure 23.1. Supply path of individual saw mills over ten years (yearly output in '000' CUM sawn timber).

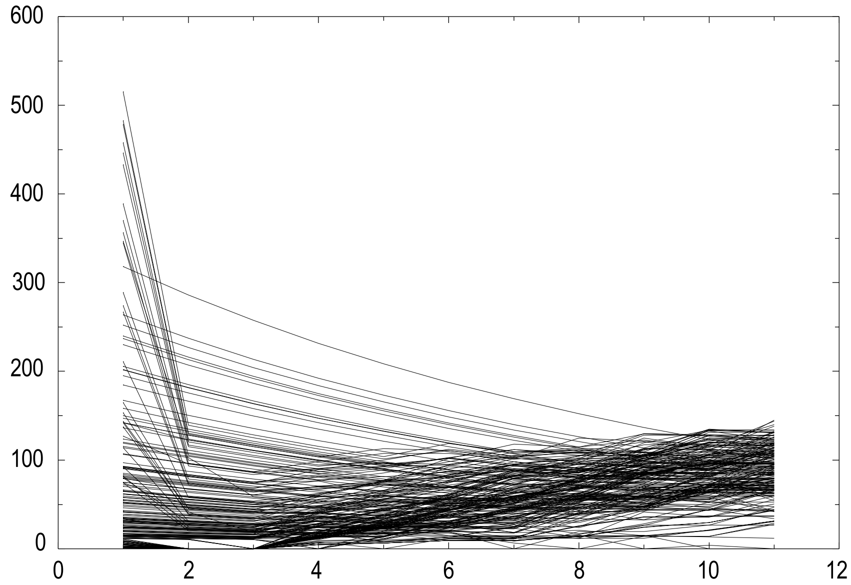


Figure 23.2. Distribution of installed capacities of individual saw mills over ten years (capacities in '000' CUM sawn timber).

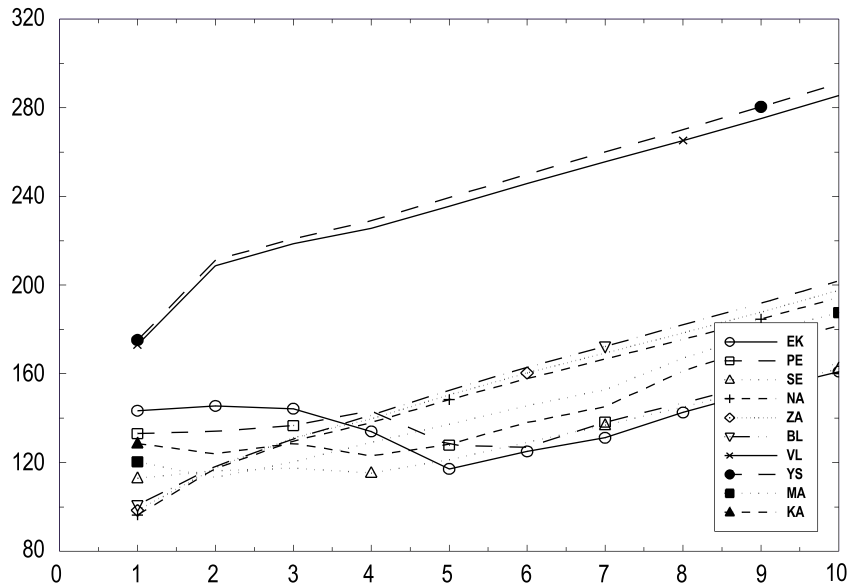


Figure 23.3. Price path for exported sawn timber over ten years in real constant (1990) US $ per CUM sawn timber. Abbreviations in the legend are: EK = Ekatherinburg, PE = Petropavlovsk, SE = Semipaladinsk, NA = Nauski, ZAS = Zabaikalsk, BL Blagovescensk, VL = Vladivostock, YS = Yuschni Sachalinsk, MA = Magadan, KA = Petropavlovsk-Kamtschtski.

observe that most of the giants are still forced to downsize in order to get closer to their optional size, which is implicitly defined by assumptions on technology parameters and resulting shapes of the cost curves. Transportation costs also appear to be of crucial importance, although the railway tariffs were assumed to be well below European levels. Some producers even had to exit the market due to their unfavorable geographic position or production scheme. Contrarily, in period four, new enterprises emerged and other enterprises began to add new capacities. The expansion of certain enterprises is due to losses of capacities of competing firms, foreign investment in the dynamic phase from period three to period four, and increased demand. After period three, the industry resized to such an extent that oversupply vanished as a market phenomenon and demand and supply began to equilibrate.

We now come to the relocation phase. Relocation can best be viewed by visual inspection of consecutive capacity plots in real geographic space such as in *Figure 23.5*. Firms depending on their competitive position have to further downsize, whereas others are able to expand. New capital additions of firms are beginning to be financed by retained capital. Analysis of the trajectories of supply and capacities

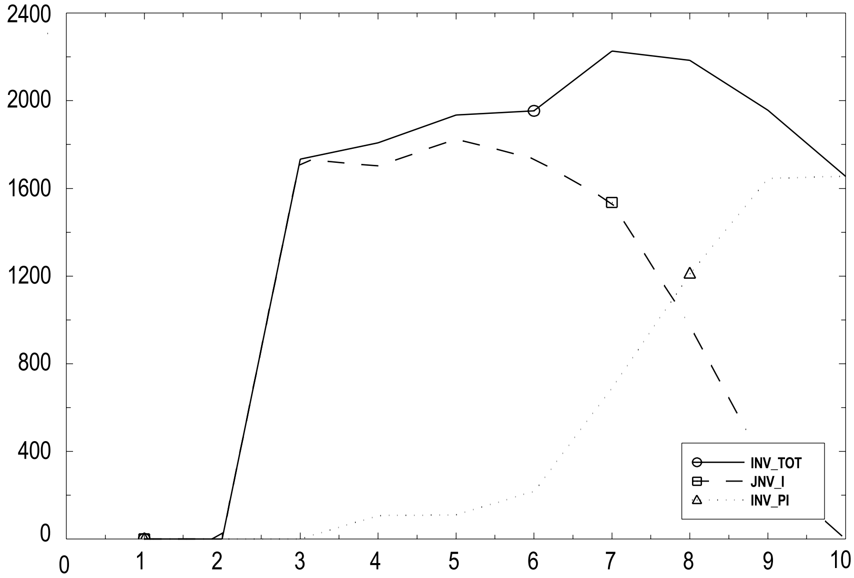


Figure 23.4. Total (inv-tot), foreign (inv-i), and investment financed from own profits (inv-pi) over ten years in '000' CUM sawn timber.

of individual saw mills reveal common patterns. This pattern is characterized by the continuous tendency of the industry to be structured around the optimal size. The optimal size, in this setting, is mainly determined by the parameters determining costs at time T. Here again, the most crucial are the indicators 'years of logging'[1] and 'terrain factor.'

23.5 GIS Representation

The option to illustrate the simulation results in a GIS representation (see *Figure 23.5*) is a very important feature of the entire modeling approach. It gives the model's user and the practitioner working for industry the opportunity to evaluate the plausibility of the results and come up with better policy conclusions. It seems that the human brain is more capable of processing visually represented data than data presented in tables or charts of more aggregated data. One develops a certain 'feeling' of what the model does if parameters change. At the same time, one learns how different policy options would affect the industry.

[1]This is the number of years a lumber mill operates in a certain area. After this period the mill is assumed to move and start operation in a new area, again with 100 percent forest cover.

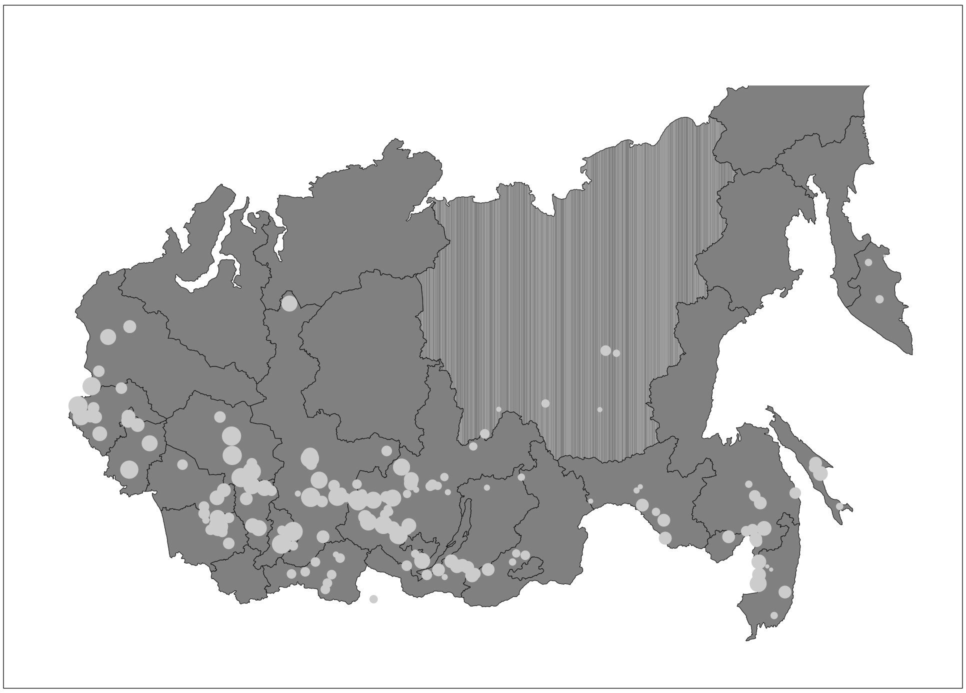


Figure 23.5. Distribution of productive capacities for lumber production in Siberia in period $t = 5$ according to the scenario discussed.

The most interesting insight gained from the GIS-representation of this scenario (*Figure 23.5*) is the rather strong concentration of enterprises in the area of West and East Siberia. It seems that this result is due to the predicted demand structure and in part due to the fact that no green-field investments were allowed in the particular set of scenarios presented. Many experts would think that on a relative scale Far Eastern regions would be able to expand more.

23.6 Summary and Conclusions

Russian industry experienced a serious economic crisis that resulted in a major drop in output in the forest industry. This caused a dramatic decrease in real employment and income in forest industry based communities. Especially in the more distant regions of Siberia the economic down-turn has been more severe compared to the Russian average. In addition, Siberia's forest sector has recently gained considerable international interest for its environmental problems. So far, little is known about the possible future path of Siberia's forest sector and its role on the international forest products markets. In this chapter a model is developed that can be used as a decision-making tool for policy analysis of various scenarios and levels of detail analyzing the forest industries of Siberia. The entire economic system is modeled on the basis of the behavior of individual firms.

The model presented here was designed to model the allocation pattern of the production of individual producers in the Siberian forest sector during the transition from a command economy to an economy based on market principles. Both the implausibility of assumptions built into standard economic models and the possibility of linking to an extensive resource database drove the modeling strategy, which was to build an easy-to-understand and easy-to-compute economic model that makes sense and makes maximum use of existing data and expert knowledge. The model, based on Vickery's Nobel-Prize-winning auction theory, simulates the possible future formation of Russia's forest industrial sector. In the model we distinguish between a static phase and a dynamic phase. The static phase describes the auction mechanism, which is applied for each period leading to a partial market clearing depending on the producer and price constraints. In the dynamic phase, product prices adjust, producers invest or depreciate capital, contract partners revise contract policies and, finally, prices change according to overall economic development.

Enormous amounts of data from many different sources and expert knowledge have been compiled and have been woven into the model. The core model is flexible in the sense that new knowledge or data can rapidly be included. This is usually achieved by combining other models with the core model. So, for example, a cost module and a demand module are attached to the core model in the current version. The cost module calculates costs as a function of a number of variables starting from forest inventory information, forest management rules, to harvesting and processing technology. The demand module predicts demand on the basis of demand functions, which were estimated with the help of econometric tools. In the current version, a Bayesian panel data model was estimated, which has never been developed in the economic literature. Another very important feature of the model is that issues of international trade can be addressed. Trade, and especially international trade flows, not only depend on predicted supply and demand patterns of importing and exporting regions, but also factors like transportation costs, loading and reloading costs, tariffs, and quotas. The effects of all these factors can be explored with the model. In addition, it is capable of modeling the effects of differences in the business approaches of different cultures on the international trade pattern. As an illustration, one can consider Chinese traders who are more concerned about prices whereas Japanese markets require high contract reliability and product quality. Another justification for the simulation approach is that the effects of wrong expectations can be quantified. So, for example, the export price to, say, Japan will fall sharply if 'too' many producers believe that Japan is their prime strategic market.

It is worth pointing out that the distinguishing feature of the entire modeling approach is that exchange is not simulated by a very specific class of trade game (i.e., general equilibrium or Bertrand competition), but by negotiations in an auction mechanism. Market power is especially important in the Russian business environment. Auctions are simulated using multiple decision criteria with different negotiation capabilities and market power of individual agents. However, optimizing behavior of any kind can be treated and implemented as special cases. In addition, the model allows for heterogeneity in the cost structure and the behavior of individual companies which can be individual producers or groups (networks) of producers.

Depending on the market power of the buyers or sellers, the algorithm either allows the buyers to purchase at a low price or the sellers to sell at a high price. Either the producers or the buyers gain relatively more from the transaction. An increasing price auction with a reserve price is iteratively conducted until either producer or buyer constraints are violated. Buyers propose contracts to a selected subset of sellers, which after a screening of all received contracts decide to accept, modify, or reject the proposals made by the buyers. Due to the nature of the auction setup, it is impossible (also theoretically by backward induction) for the individual agents to compute their optimal strategy in the auction nor is it possible to compute the optimal strategy over periods using an inter-temporal optimization procedure. Nonetheless, the agents' behavior tends towards a quasi-optimal state in a distant future period.

Typically, output is given at an individual enterprise level of the final product. Trade flows between individual agents, prices negotiated by individual bargain, supply slacks, capital formation, profits, investment, and many other details can be reported on an individual mill or even contract level. Despite the fact that the model has not been finely tuned, preliminary results seem to indicate that it is capable of matching reasonably well the pattern of output decline and relocation. Also price development, especially for export prices, can almost realistically be reconstructed, although the quality of these data can vary considerably.

These are all very encouraging results. Nonetheless, the model will still have to be further extended and finely tuned. The most recent results from theory will have to be continuously incorporated into the model. There is especially a need to further strengthen the theoretic basis for the transition algorithms used in the dynamic phase. Also, empirical experiments will have to be carried out in order to pin down negotiation behavior across different cultures. The biggest problem remaining is the correct parameterization of the model. As already stated, it is possible to finely tune the model in such a way that it can redraw the actual allocation pattern of production or price developments. However, just as optimization procedures can be trapped in local minima or maxima, the parameter constellation (and there are many parameters to be tuned) can accidentally lead to a good match with the actual reference data, but not reflect the truth. There is still an automated procedure to be developed that could find the 'true' parameters with reasonable computational costs.

References

Backman, Ch.A., 1996, The Russian forest sector: Prospects for trade with the former Soviet republics, *Post Soviet Geography and Economics*, **37**(1):16–59.

Baltagi, B.H., 1995, *Econometric Analysis of Panel Data*, John Wiley and Sons Ltd., New York, NY, USA.

Gaspar, J., and Glaeser, E., 1996, Information Technology and the Future of Cities, NBER Working Paper No. W5562, National Bureau for Economic Research, Cambridge, MA, USA.

Hamilton, J.H., 1994, *Time Series Analysis*, Princeton University Press, Princeton, NJ, USA.

Hanson, G.M., 1998, North American economic integration and industry location, *Oxford Review of Economic Policy*, **14**:30–44.

Maddala, G.S., 1993, *The Econometrics of Panel Data*, Edward Elgar Publishing, Cheltenham, UK.

Obersteiner, M., 2000, Restructuring, Efficiency and Output Decline of Russian Industries and Regions, Interim Report IR-99-066, International Institute for Applied Systems Analysis, Laxenburg, Austria, available at: http://www.iiasa.ac.at/Publications/Catalog/PUB_PROJECT_FOR.html

Obersteiner, M., 1998, The Pan Siberian Forest Industry Model (PSFIM): A theoretical concept for forest industry analysis. Interim Report IR-98-033, International Institute for Applied Systems Analysis, Laxenburg, Austria, available at: http://www.iiasa.ac.at/Publications/Catalog/PUB_PROJECT_FOR.html

Voevoda, I.N., 1985, *Lesnaja i lesooprabatybajushaja promyshlenost Sibiri*, Nauka, Institute for Economics and Industrial Organization, Novosibirsk, Russia (in Russian).

Affiliation of Authors

Masao Arakawa, Department of Reliability-Based Information System Engineering, Faculty of Engineering, Kagawa University, Kagawa, Japan

Masaaki Asada, Mitsubishi, Chemical Corporation Mizushima Plant, Kurashiki, Japan

Thomas Barth, Department of Information Systems, University of Siegen, Siegen, Germany; barth@fb5.uni-siegen.de

Paweł Bartoszczuk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland; bartosz@ibspan.waw.pl

Wieslaw Ciechanowicz, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland; ciechano@ibspan.waw.pl

Graham Cox, National Audubon Society of New York State, Albany, NY, USA; gcox@audubon.org

Richard Cumpston, Cumpston Sarjeant Pty Ltd., Melbourne, Australia; rich ard_cumpston@cumsar.com.au

Jon Erickson, Rensselaer Polytechnic Institute, Sage Lab, Troy, NY, USA; erickj @rpi.edu

Günther Fischer, International Institute for Applied Systems Analysis, Laxenburg, Austria; fisher@iiasa.ac.at

Matti Flinkman, Department of Forest Management and Products, Swedish University of Agricultural Sciences, Uppsala, Sweden and Department of Statistics, Umeå School of Business and Economics, Umeå University, Umeå, Sweden; Matti.Flinkman@stat.umu.se

Manfred Grauer, Department of Information Systems, University of Siegen, Siegen, Germany; grauer@fb5.uni-siegen.de

Kiyotada Hayashi, National Agricultural Research Center (Tohoku), Iwate, Japan; hayashi@affrc.go.jp

Pirja Heiskanen, Fortum Power and Heat Oy, Fortum, Finland; pirja.heiskanen@ fortum.com

Klaus Hubacek, Rensselaer Polytechnic Institute, Troy, NY, USA; huback@rpi. edu

Hiroshi Ishikawa, Department of Reliability-Based Information System Engineering, Kagawa University, Kagawa, Japan

Stefan Kaden, Wasy Institute for Water Resources, Planning and Systems Research Ltd., Berlin, Germany; S.Kaden@wasy.de

Pekka Korhonen, Department of Economics and Management Science, Helsinki School of Economics, Helsinki, Finland; Pekka.Korhonen@Hkkk.fi

Mikulas Luptacik, Department of Economics, Vienna University of Economics and Business Administration, Vienna, Austria; mikulas.luptacik@wu-wien.ac.at

Marek Makowski, International Institute for Applied Systems Analysis, Laxenburg, Austria; marek@iiasa.ac.at

Wojtek Michałowski, Faculty of Administration, University of Ottawa, Ottawa, Canada; wojtek@admin.uottawa.ca

Ingo Michels, Wasy Institute for Water Resources, Planning and Systems Research Ltd., Berlin, Germany; I.Michels@wasy.de

Yasuo Mitani, Faculty of Engineering, Fukuyama University, Fukuyama, Japan; mitani@fuee.fukuyama-u.ac.jp

AbuBakr AbdelAziz Mohamed, Crop, Soil and Water Sciences Division, International Rice Research Institute, Manila, Philippines; mabubakr@cgiar.org

Noboru Nakasako, School of Biology-Oriented Science and Technology, Kinki University, Wakayama, Japan; nakasako@info.waka.kindai.ac.jp

Hirotaka Nakayama, Department of Applied Mathematics, Konan University, Kobe, Japan; nakayama@konan-u.ac.jp

Sten Nilsson, International Institute for Applied Systems Analysis, Laxenburg, Austria; nilsson@iiasa.ac.at

Michael Obersteiner, International Institute for Applied Systems Analysis, Laxenburg, Austria; oberstei@iiasa.ac.at, and Institute for Advanced Studies, Vienna, Austria; oberstei@ihs.ac.at

Jolanta Pakulska, Research Institute of Economic Development, Warsaw School of Economics, Warsaw, Poland; jpakul@sgh.waw.pl

Zdzisław Pawlak, University of Information Technology and Management, Warsaw, Poland; zpw@ii.pw.edu.pl

Jan Sendzimir, International Institute for Applied Systems Analysis, Laxenburg, Austria; sendzim@iiasa.ac.at

Roman Słowiński, Institute of Computing Science, Poznan University of Technology, Poznan, Poland

Sigrid Stagl, Department of Economics, Vienna University of Economics and Business Administration, Vienna, Austria; stagl@wu-wien.ac.at

Laixiang Sun, International Institute for Applied Systems Analysis, Laxenburg, Austria; sun@iiasa.ac.at

Yoshitaka Takahashi, Department of Livestock and Grassland Science, National Agricultural Research Center for Western Region, Shimane, Japan; yositaka@anim.affrc.go.jp

Masahiro Tanaka, Konan University, Kobe, Japan; m_tanaka@konan-u.ac.jp

Marites M. Tiongco, Social Sciences Division, International Rice Research Institute, Manila, Philippines; m.tiongco@cgiar.org

Eiji Watanabe, Department of Information Systems and Management Science, Konan University, Kobe, Japan; watanabe@we-sv.mis.konan-u.ac.jp

Jaap Wessels, EURANDOM, Eindhoven, The Netherlands; wessels@win.tue.nl

David Wiberg, International Institute for Applied Systems Analysis, Laxenburg, Austria; wiberg@iiasa.ac.at

Szymon Wilk, Institute of Computing Science, Poznan University of Technology, Poznan, Poland; szymon.wilk@cs.put.poznan.pl

Ye Boon Yun, Department of Reliability-Based Information System Engineering, Kagawa University, Kagawa, Japan; yun@eng.kagawa-u.ac.jp