DISSERTATIONS IN

# FORESTRY AND NATURAL SCIENCES

HARRI NISKA

# *Predictive Data-Driven Modeling Approaches in Environmental Management Decision-Making*

PUBLICATIONS OF THE UNIVERSITY OF EASTERN FINLAND
*Dissertations in Forestry and Natural Sciences*

UNIVERSITY OF
EASTERN FINLAND

**HARRI NISKA**

# *Predictive Data-Driven Modeling Approaches in Environmental Management Decision-Making*

Publications of The University of Eastern Finland
Dissertations in Forestry and Natural Sciences
No 60

Academic Dissertation
To be presented by permission of the Faculty of Natural Sciences and Forestry for public examination in the Auditorium L21 at the University of Eastern Finland, Kuopio, on January, 20th, 2012, at 12 o'clock noon

Department of Environmental Science

Kopijyvä
Kuopio, 2011
Editor: Research Director Pertti Pasanen
Lecturer Sinikka Parkkinen, Prof. Pekka Kilpeläinen
Distribution:
Eastern Finland University Library / Sales of Publications
P.O.Box 107, FI-80101 Joensuu, Finland
tel. +358-50-3058396
http://www.uef.fi/kirjasto
Front cover picture:
Vuorela, Siilinjärvi, by the Author

ISBN: 978-952-61-0646-5
ISBN: 978-952-61-0647-2 (PDF)
ISSNL: 1798-5668
ISSN: 1798-5668
ISSN: 1798-5676 (PDF)

Author's address: Department of Environmental Science
University of Eastern Finland
P.O.Box 1627
70211 Kuopio
FINLAND
email: harri.niska@uef.fi

Supervisors: Professor Mikko Kolehmainen, D.Sc. (Tech.)
University of Eastern Finland
Department of Environmental Science
P.O.Box 1627
70211 KUOPIO
FINLAND
email: mikko.kolehmainen@uef.fi

Professor Emeritus Juhani Ruuskanen, PhD.
University of Eastern Finland
Department of Environmental Science
P.O.Box 1627
70211 KUOPIO
FINLAND
email: juhani.ruuskanen@uef.fi

Research Professor Jaakko Kukkonen, PhD.
Finnish Meteorological Institute
Air Quality Research
P.O.Box 503
00101 HELSINKI
FINLAND
email: jaakko.kukkonen@fmi.fi

Docent Ari Karppinen, D.Sc. (Tech.)
Finnish Meteorological Institute
Air Quality Research
P.O.Box 503
00101 HELSINKI
FINLAND
email: ari.karppinen@fmi.fi

Reviewers:

Professor Enso Ikonen, D.Sc. (Tech)
University of Oulu
Department of Process and Environmental Engineering
P.O.Box 4300
90014 OULU
FINLAND
emal: enso.ikonen@oulu.fi

Docent Francesco Corona, D.Sc. (Tech)
Aalto University
Department of Information and Computer Science
P.O.Box 15400
00076 AALTO
FINLAND
email: francesco.corona@aalto.fi

Opponent:

Professor Ari Jolma, D.Sc. (Tech)
Aalto University
Department of Civil and Environmental Engineering
P.O.Box 12100
00076 AALTO
FINLAND
email: ari.jolma@aalto.fi

Niska, Harri
Predictive data-driven modeling approaches in environmental management decision-making
University of Eastern Finland, Department of Environmental Science, 2011
Publications of the University of Eastern Finland, Dissertations in Forestry and Natural Sciences, No 60
ISBN: 978-952-61-0646-5
ISBN: 978-952-61-0647-2 (PDF)
ISSNL: 1798-5668
ISSN: 1798-5668
ISSN: 1798-5676 (PDF)

## ABSTRACT

Computational data-driven models are increasingly required to support conclusions and to aid in reaching in-time and sufficient decisions in environmental research, planning and management. The aim of this thesis was to evaluate the usability of modern computational methods and related data-driven modeling (DDM) schemes for solving the predictive modeling problems associated with environmental management decision-making. The selected case studies included were (i) the forecasting of urban airborne pollutant concentrations, (ii) the characterization of physicochemical and biological properties of chemical substances, using quantitative structure activity relationships (QSARs) and chemical grouping, as well as (iii) the prediction of species-specific forest attributes using airborne laser scanning (ALS) data.

First, a brief overview into the domain of application and the modeling problems to be studied is given. There follows an introduction to the computational data-driven modeling approaches, including the main modeling methods used in this thesis, namely multi-layer perceptron (MLP), support vector regression (SVR), self-organizing map (SOM) and Sammon's mapping. Predictive modeling approaches, based on the selected modeling methods, are then evaluated and discussed, with specific conclusions in each application domain. Finally, the significance of the work is assessed and recommendations for future work are laid out.

The results of the air quality studies show that the MLP network yields moderately good general performance for the prediction of airborne pollutant concentrations of $NO_2$ and $PM_{2.5}$. It is also shown that the performance of MLP network can be enhanced in operational urban air pollution forecasting, using the predictions of numerical weather prediction (NWP) model as input. The performance of the MLP network is, however, obtained to be degenerated in the course of peak pollution episodes. Further, the results obtained show that SOM

and MLP are appropriate methods for recovering incomplete air quality datasets to complete form required by the modeling. In chemical modeling studies, Sammon's mapping and its combination with regression-based QSAR models is shown to be a powerful approach for discovering and visualizing chemical substance groups. Such a chemical grouping approach could be used as an alternative for conventional, laboratory-based testing strategies in REACH (2006/1907/EC) when characterizing and predicting unknown physicochemical properties and environmental and health effects of target chemical substances. Lastly, the results of ALS-based forest inventory studies show that MLP and SVR produce reliable estimates for species-specific forest attributes, which are increasingly needed by forest management and energy production applications. The performance of MLP and SVR is found to be comparable to the corresponding performance of current ALS-based forest inventory methods.

In addition to the novel applications of the modeling methods, the main innovation of this thesis was to show the usability of GA-based optimization schemes for selecting the appropriate structure of air quality and forest inventory models. Even though approaches based on the use of GA have been presented in the related fields of environmental modeling, they have not been previously applied in the selected application domains to this extent.

The results and observations of this thesis in general suggest that the computational methods studied are well-suited for solving complex predictive modeling problems in environmental management. In the future, further development of the modeling is required, especially, in respect to the modeling and prediction of rare and spatially dependent processes. A combination of the modern data-driven modeling methods and geostatistical modeling methods is thus one potential research direction. In addition, more emphasis should be placed on improving the mechanistic interpretation of the models in order to improve their (regulatory) acceptance. This requires the development of hybrid modeling approaches, where physical information about underlaying system is encapsulated at some level into the data-driven modeling.

Universal Decimal Classification: 004.9, 502.14, 502.3, 630*5
CAB Thesaurus: information systems; computer techniques; models; neural networks; data processing; optimization; prediction; environmental assessment; environmental management; decision making; air quality; chemicals; risk assessment; structure activity relationships; forest inventories; remote sensing; aerial surveys
Yleinen suomalainen asiasanasto: informatiikka; tietojärjestelmät; tiedonlouhinta; mallintaminen; mallit; optimointi; neuroverkot; geneettiset algoritmit; ympäristöongelmat; päätöksenteko; ennusteet; ilmanlaatu; kemikaalit; riskinarviointi; metsänarviointi; kaukokartoitus; laserkeilaus

# *Acknowledgements*

The work leading to this thesis has been carried out in several research projects at the Department of Environmental Science, University of Eastern Finland (formerly University of Kuopio), during the years 2002–2010. The work has been financially supported by the European Union (APPETISE EU project, IST-99-11764), the Academy of Finland (FORECAST, the project no. 49946), Neste Oil Corporation, the Ministry of Education and the University of Eastern Finland (the project, Dnro. 3486/11/07).

First of all, I would like to express my gratitude to my supervisors Prof. Mikko Kolehmainen and Prof. Emeritus Juhani Ruuskanen (University of Eastern Finland). Without their constant support and open minded brainstorming, this work would not have been possible. I am also deeply grateful to my other supervisors Prof. Jaakko Kukkonen and Docent Ari Karppinen (Finnish Meteorological Institute) for all their guidance and valuable comments during this work.

I would like to thank my co-authors Teri Hiltunen, Docent Kari Tuppurainen, Jukka-Pekka Skön, Heikki Junninen, Minna Rantamäki (Finnish Meterological Institute), Prof. Timo Tokola, Prof. Matti Maltamo, Docent Petteri Packalén and Dr. Anthony K. Mallett (Experien Health Sciences Ltd) for the fruitful collaboration and their valuable contribution during this work. I am also indebted to many of my colleagues, especially Dr. Teemu Räsänen, Dr. Jarkko Tissari, Dr. Mauno Rönkkö, Dr. Hannu Poutiainen, Kari Pasanen, Jarkko Tiirikainen, Mikko Heikkinen, Juha Parviainen, Mika Raatikainen, Tuomas Huopana, Jukka Saarenpää, Okko Kauhanen, Markus Stocker, Xavier Albacete, for all support during this study. I would also like to thank Dr. Dimitris Voukantsis and Prof. Kostas Karatzas (Aristotetle University of Thessaloniki) for their long-term collaboration and, especially, for the possibility for carrying out part of this work in their research group in Thessaloniki.

I am indebted to the pre-examiners of this thesis, Prof. Enso Ikonen (University of Oulu) and Docent Fransesco Corona (Aalto University), for their constructive feedback and suggestions, which substantially helped to enhance the structure of the thesis, as well as several details in it. I am also deeply grateful to Karin Koivisto for the proofreading of the manuscript and all the practical help in the process. I would also like to thank the personnel of the Department of Environmental Science, and especially Marja-Leena Patronen, Ritva Karhunen and Kaija Ahonen for their valuable support in significant project-related details covering all sorts of practical issues.

Special thanks go to my parents Riitta and Mauri. My mother Riitta was actually the person who guided me to seek this discipline in the first place. Most of all, I wish to express my warmest thanks to my wife Jaana, and to our children Sofia, Rasmus, Veikka and Vilma, for their love and constant understanding during this process.

Kuopio, December 2011

Harri Niska

# *List of publications*

This thesis is based on the following original publications referred to in the text by their Roman numerals (**Papers I–V**):

**Paper I** Junninen H., Niska H., Tuppurainen K., Ruuskanen J., Kolehmainen M. (2004) Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895–2907.

**Paper II** Niska H., Hiltunen T., Karppinen A., Ruuskanen J., Kolehmainen M. (2004) Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence 17, 159–167.

**Paper III** Niska, H., Rantamäki, M., Hiltunen, T., Karppinen, A., Kukkonen, J., Ruuskanen, J., Kolehmainen, M. (2005) Evaluation of an integrated modeling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. Atmospheric Environment 39, 6524–6536.

**Paper IV** Niska, H., Tuppurainen, K., Skön, J.-P., Kolehmainen, M., Mallett, A.K. (2008) Characterization of the chemical and biological properties of molecules with QSAR/QSPR and chemical grouping, and its application to a group of alkyl ethers. SAR and QSAR in Environmental Research 19, 263–284.

**Paper V** Niska, H., Skön, J.-P., Packalén, P., Tokola, T., Maltamo, M., Kolehmainen, M. (2010) Neural networks for the prediction of species-specific stem volumes using airborne laser scanning and aerial photographs. IEEE Transactions on Geoscience and Remote Sensing 48, 1076–1085.

# *The author's contribution*

The publications of this thesis have originated from several research projects carried out mainly in the Department of Environmental Science, University of Eastern Finland (formerly the University of Kuopio). The author has had a significant role in each paper, the contribution varying from case to case, as explained below.

In **Paper I**, the author selected the methods, designed and implemented the modeling schemes, and conducted the modeling experiments jointly with H. Junninen. The author was responsible for designing and performing the experimental comparison of the methods. The author analyzed the results jointly with the co-authors and was responsible for writing of the paper. The role of K. Tuppurainen, J. Ruuskanen and M. Kolehmainen was supervisory.

In **Papers II and III**, the author was responsible for selecting, innovating and implementing the modeling schemes. The author designed and conducted the model computations and the statistical evaluation of the methods. The author analyzed the results jointly with the co-authors and was the principal writer of the papers. The role of T. Hiltunen was to aid with the practical implementation of the MLP-based modeling schemes. The role of M. Rantamäki was to offer her expert knowledge in NWP and to aid with the processing of NWP data. The work was done under the supervision of M. Kolehmainen, J. Ruuskanen, A. Karppinen and J. Kukkonen.

In **Paper IV**, the author designed and performed modeling experiments, and analyzed the modeling results jointly with the co-authors. The author was responsible for writing the paper together with K. Tuppurainen. The role of J.P. Skön was to perform the QSAR computations. The role of K. Tuppurainen was supervisory, including help in QSARs and in performing QSAR computations and chemical grouping. The role of A.K. Mallett was to offer his expert knowledge in the chemical risk assessment and to help in improving the final quality of the paper. The role of M. Kolehmainen was supervisory.

In **Paper V**, the author was responsible for selecting and implementing the modeling schemes and conducting the modeling experiments. The author interpreted the results jointly with the co-authors and was the principal writer of the paper. The role of J.P. Skön was to help in ALS data processing and model specification. The roles of P. Packalén, T. Tokola and M. Maltamo were to offer their expert knowledge in the ALS-based forest inventory methods and ALS data processing. The role of M. Kolehmainen was supervisory.

# *Contents*

**ABBREVIATONS**

| | |
|---|---|
| ANN | Artificial Neural Network |
| ALS | Airborne Laser Scanning |
| AQF | Air Quality Forecasting |
| CI | Computational Intelligence |
| DDM | Data-Driven Model |
| DET | Deterministic Dispersion Model |
| EA | Evolutionary Algorithm |
| EC | European Commission |
| ECHA | European Chemicals Agency |
| EPA | US Environmental Protection Agency |
| EU | European Union |
| GA | Genetic Algorithm |
| HIRLAM | High Resolution Limited Area Model |
| MLR | Multiple Linear Regression |
| KDD | Knowledge Discovery in Databases |
| LOO | Leave-One-Out Cross validation |
| LR | Linear Regression |
| LS | Method of Least Squares |
| MLP | Multi-Layer Perceptron |
| MLR | Multiple Linear Regression |
| MSN | Most Similar Neighbor |
| MOGA | Multi-Objective Genetic Algorithm |
| NN | Nearest Neighbor Regression |
| NWP | Numerical Weather Prediction |
| $NO_2$ | Nitrogen Dioxide |
| $PM_{10}$ | Particles smaller than 10 μm in aerodynamic diameter |
| $PM_{2.5}$ | Particles smaller than 2.5 μm in aerodynamic diameter |
| RS | Remote Sensing |
| OECD | Organization for Economic Co-operation and Development |
| PCA | Principal Component Analysis |
| PLS | Partial Least Squares |
| REACH | Registration, Evaluation and Authorization of Chemicals |
| SA | Sensitivity Analysis |
| SAR | Structure Activity Relationship |
| SOM | Self-Organizing Map |
| SVR | Support Vector Regression |
| QSAR | Quantitative Structure Activity Relationship |
| QSPR | Quantitative Structure Property Relationship |

**LIST OF FREQUENTLY USED SYMBOLS**

| | |
|---|---|
| $b$ | bias |
| $\beta$ | regression coefficient |
| $e$ | error residual |
| $f$ | function |
| $i, j$ | general indices |
| $n$ | number of input vectors (data lines) |
| $p$ | number of dimensions (variables) |
| $t$ | time |
| $\mathbf{w}$ | weight vector |
| $\mathbf{X}$ | input data matrix |
| $\mathbf{Y}$ | target data matrix |
| $\mathbf{x}$ | input vector |
| $x$ | observed input variable |
| $\mathbf{y}$ | target data vector |
| $y$ | observed target variable |
| $\hat{y}$ | estimated target variable |

# 1 Introduction

Nowadays, there is an increasing need for powerful and reliable computational models that can be used to support decision-makers in managing and regulating environmental issues. Concerning urban air pollution control, for instance, the reliable site-specific air quality forecasts are required in order to set up emergency response plans and potential practical measures such as traffic limitations during peak pollution situations. In chemical risk assessment, computational non-testing (in-silico) methods are required as alternatives for conventional laborious and expensive in-vivo and in-vitro testing strategies. In natural resource management, respectively, powerful remote sensing (RS) methods are an essential part of decision support and information systems and are increasingly required to replace time-consuming field-assessment procedures.

The environment is a highly complex system, characterized with ill-defined natural and anthropogenic interactions and feed-back loops between systems (e.g. Green and Klomp, 1998; Haykin and Principe, 1998). In addition, processes themselves usually have an inner structure, where different parts of the process influence each other with delays (Haykin and Principe, 1998). Atmospheric pollutants, for instance, are a consequence of natural and anthropogenic (e.g. traffic and industry) emission processes, chemical reactions between species, solar radiation, temperature and other interactive processes (San José et al., 2009). According to Green and Klomp (1998) the complexity of environmental systems can be characterized by spatial and temporal scales, non-linear interactions and feedback loops, high number of influencing factors, criticality and human influence.

The complexity of environment poses many challenges in modeling (e.g. Green and Klomp, 1998). Limitations are associated especially with the analysis of complex and ill-defined systems, such as biological systems, weather-related phenomena, fluid turbulence and radar backscatter from the sea surface, characterized through massive interactions among different parts of a system or nonlinear phenomena (Haykin and Principe, 1998). In such conditions physical (mechanistic) based modeling usually fails and statistical approaches are required to establish unknown relationships from the data (e.g. Gardner and Dorling, 1998). The standard statistical approaches are, however, in many cases relatively unsophisticated for dealing with environmental data, characterized with missing data, noisy or collinear variables (e.g. McCune, 1997), heterogeneous distributions (e.g. Rong, 2000), and non-linear, time-delayed interconnections between variables (e.g. Gardner and Dorling, 1998).

In recent years data-driven modeling (DDM), which rely on the methods of computational intelligence (CI), have been increasingly adopted for solving complex modeling problems in environmental sciences (e.g. Krasnopolsky and Chevallier, 2003; Solomatine, 2005; Haupt et al., 2008). Basically, DDM can be regarded as a general framework for the data-based (empirical) modeling, having a limited knowledge about the physical behavior of the system (Solomatine and Ostfeld, 2008). The DDM approach can thus in principle act as a complementary to physical modeling, which is based on the incorporation of known physical, chemical or biological mechanisms in the modeling.

The objective of this thesis was to evaluate the usability of the modern computational methods and related DDM approaches for solving complex predictive modeling problems associated with environmental management decision-making. The selected case studies were:

- Forecasting of urban airborne pollutant concentrations
- Characterization of physicochemical and biological properties of chemical substances, using quantitative structure activity relationships (QSARs) and chemical grouping
- Predicting species-specific forest attributes using airborne laser scanning (ALS) data

The study was carried out through experimental model design and evaluation work with an examination of the external validity using a comparison of model output with the experimental data. In each application domain suitable experimental data were available, as well as experts in the field who could participate in guiding the work.

The selected computational methods contained up-to-date artificial neural networks (ANNs) and related methods, namely multi-layer perceptron (MLP), support vector regression (SVR), self-organizing map (SOM) and Sammon's mapping, all of which have been previously shown to exhibit good processing and modeling capability for the environmental data (e.g. Canu and Rakotomamonjy, 2001; Lu et al., 2002; Kolehmainen, 2004; Lu and Wang, 2005). Moreover, the methods were compared with other modeling approaches previously adopted in the application/problem domains. In addition to the novel applications of the modeling methods, the main innovation of the thesis was to show the usability of GA-based optimization schemes for selecting appropriate input variables and structure of the data-driven models. Even though approaches based on the use of GA have been presented in the related fields of environmental modeling, they have not been previously applied in the selected application areas to this extent.

In the studies with air quality forecasting (AQF), the main objective was to investigate the usability of MLP-based modeling for forecasting urban airborne pollutant concentrations of nitrogen dioxide ($NO_2$) and particular matter ($PM_{2.5}$)

and, particularly, to examine the accuracy of the modeling in course of infrequent occurrence of peak pollution episodes. In addition, the objective of air quality studies was to investigate the accuracy of MLP in an operational condition, where numerical weather prediction (NWP) data are available for the modeling. This is important since the evaluation of ANN-based air quality models has been mainly based on the use of meteorological measurement data instead of actual NWP predictions (e.g. Kolehmainen et al., 2001; Kukkonen et al., 2003). Further, novel computational approaches for imputing missing data in air quality datasets were examined in order to recover incomplete data to the complete format required by the MLP- based modeling schemes.

In the studies with QSARs and chemical grouping, the objective was to investigate how the methods can be used to characterize unknown physicochemical properties, and environmental and health effects of target chemical substances within the framework of the EU's REACH regulation (2006/1907/EC). The principal focus was on the discovery of chemical groups for a set of target chemical substances and a set of reference chemical substances in respect to the calculated molecular descriptor data. In REACH, the information derived from chemical grouping can allow the use of the so-called read-across approach, where unknown properties of target chemical substances are interpolated, based on the existing data of reference chemical substances belonging to the same chemical group. Such read-across approaches are expected to be necessary when reducing the need of the extensive laboratory-based testing procedures often based on the use of animal experiments.

Lastly, the studies in the field of forest inventory, aimed at evaluating the accuracy of the ANN methods, namely MLP, SVR and SOM in the prediction of species-specific forest attributes using ALS data. Previously, the ALS-based forest inventory models have been based largely on the use of conventional regression methods, and so far ANNs have not been examined in this field to this extent. The ANN methods were compared to the non-parametric *k*-most similar neighbor (*k*-MSN), which has recently been adopted in the ALS-based forest inventory studies (e.g., Packalén, 2010).

The structure of the thesis is divided as follows. First, a brief introduction into the domain of application and the modeling problems to be studied is given (Chapter 2). In each domain, general research hypotheses and potential solutions are proposed. There follows the presentation of the modern computational approaches for the processing of environmental data, including the main modeling methods studied in this thesis (Chapter 3). Next, the aims of the thesis are briefly summarized, followed by the presentation of material and methods and the evaluation of the key results and findings in each case studies (Chapter 4). Finally, the significance of the work is assessed and recommendations for future work are laid out (Chapter 5).

# 2 The domain of application

## 2.1 ENVIRONMENTAL INFORMATICS

The work presented here falls into the discipline of environmental informatics. Environmental informatics is a branch of applied computer science, which develops and uses information technology and computational methods for environmental protection, research and engineering (e.g. Avouris and Page, 1995; Page and Hilty, 1995; Green and Klomp, 1998; Kolehmainen, 2004). According to Page and Hilty (2005), environmental informatics can be defined as follows:

*"Environmental informatics is a special sub-discipline of Applied Informatics dealing with the methods and tools of computer sciences for analyzing, supporting and setting up those information processing procedures which are contributing to the investigation, removal and minimization of environmental burden and damages."*

On the other hand, the role of environmental informatics can be seen as a mediator between environmental sciences and modern information technology, providing novel data-driven solutions based on processing collected data into the information and knowledge needed for problem solving in the domain (Figure 1).

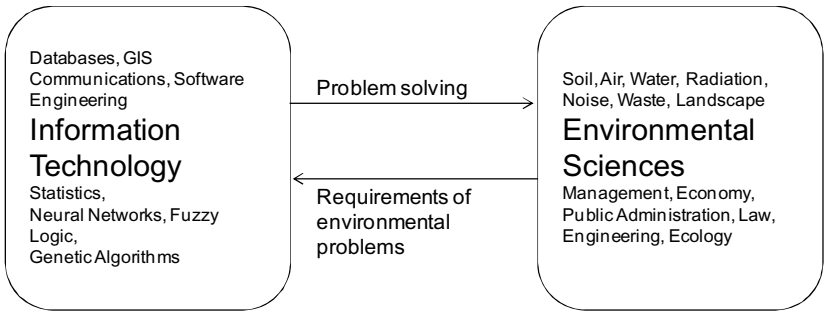


Figure 1. The role of environmental informatics between environmental sciences and information technology (modified from Page and Rautenstrauch, 2001).

Such a problem solving requires developing and studying adequate methods for effective processing of environmental data (e.g. Avouris and Page, 1995; Kolehmainen, 2004). From this point of view, it is relevant to pay an attention to the complexity of environmental systems/problems and its various appearances in collected environmental data, which pose many challenges for the problem solving.

## 2.2 CHALLENGES WITH ENVIRONMENTAL DATA

Green and Klomp (1998) have arranged the sources of the complexity of environmental systems into the following categories:

- Spatial and temporal scales
- Non-linear interactions and feedback loops
- High number of influencing factors
- Human influence

Data produced from such "non-controllable" environment represent a combination of several processes of multivariate origin, which render non-linear, chaotic and noisy characteristics of nature.

First, the multitude of variables are required to be collected to achieve sufficient representation on influencing spatiotemporal factors, their possible interactions and feedback loops as well as human influence in the modeling. Frequently, the base dataset have been fused with external dataset from the same geographic region, which is expected to increase the amount of information on the underlaying problem for the modeling/analysis. These requirements results in large and heterogeneous data matrices, with different spatial and temporal scales, different dimensions, modes and orders. Furthermore, data matrices produced are often correlated or may contain inner structures, where different variables are interconnected each other with non-linearity and delays. An important aspect is also seasonality, i.e., primary factors to be analyzed and modeled are originated from cycles of nature and human activity. Consequently, several years of measurement data are required to be collected in order to capture all relevant conditions, changes and trends in underlaying systems.

In addition, environmental data are influenced significantly by deficiencies and errors in the data collection, which should be considered when developing and studying the methods (e.g. Cherkassky et al., 2006; Barry and Elith, 2006). A problem encountered most frequently is missing data, which is due to device failures, human errors or insufficient sampling and spatial coverage. Other common problems include measurement errors, noise and outliers, which are due to errors by devices or human operators and erroneous calibration of devices.

## 2.3 ENVIRONMENTAL MANAGEMENT DECISION-MAKING

The methods studied in Environmental Informatics are often associated with information processing procedures of environmental management decision-making. Environmental management is a broad concept, but basically it can be seen as the process for managing and controlling of human activities and their impacts on the environment. The main components of environmental management and its relation to Environmental Informatics are depicted in Figure 2.

The role of Environmental Informatics in environmental management is mainly in studying appropriate methods for collecting, retrieving, storing and processing measurement data into useful information and knowledge needed by the decision-maker. An essential component of that process is the modeling. Basically, the modeling is required to aid in reaching in-time and sufficient management decisions e.g. by means of replacing laborious and expensive measurement procedures, filling in information gaps of monitoring and producing new information on complex environmental systems (e.g. Avouris and Page, 1995; Maier et al., 2008). Usually the objective of the modeling is to predict unknown properties, effects or events of environmental systems from the basis of the available physical information and/or collected measurement data from the underlaying system.

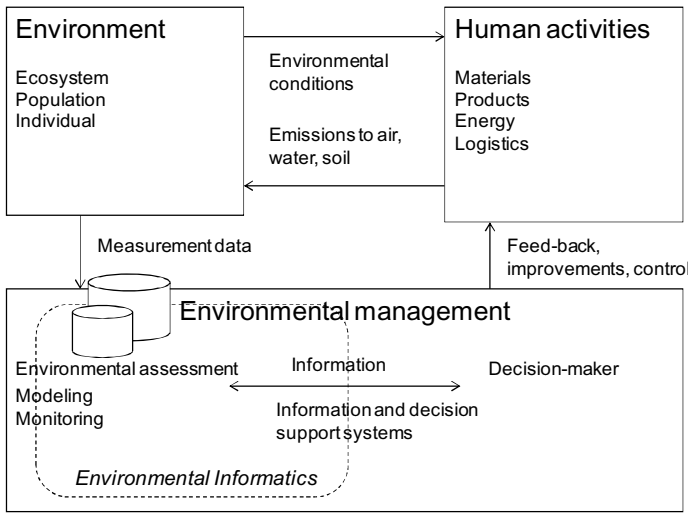


*Figure 2. Main components of environmental management decision-making and relation to Environmental Informatics.*

Jakeman et al. (2006) have demonstrated the importance of models in environmental management options and separated models into different model families, which include e.g. empirical, data-based, statistical models, process-based models (called deterministic models), agent-based models and rule-based models. Following Seppelt (2003), it is useful to divide models into the following model categories:

- Theory-based (white-box) models
- Empirical (black-box) models
- Theory-influenced empirical (grey-box) models

In complex and ill-defined situations, as studied in this thesis, variables characterizing the behavior of the system can be measured and used to construct a data-based model. Such empirical data-based modeling is a natural choice, since theory-based (physical) modeling suffers the lack of prior-knowledge on underlaying physical laws and relationships of the system. In addition, theory-based modeling may be time-consuming or may lead to unnecessarily complex models.

The modeling is required in various fields of environmental management, which covers for instance air quality, climate change, wastes and chemicals and natural resources. In this thesis, the computational methods are studied in the fields associated to urban air pollution control, chemical risk assessment and management and natural resources management. Next, the domains of the application and the modeling problems studied are briefly introduced.

## 2.4 URBAN AIR QUALITY CONTROL

Urban air quality (AQ) has emerged as an acute environmental problem, especially for densely populated metropolitan areas, causing negative effects on health, ecosystems and materials. To prevent further decline in air quality it is necessary (Kolehmainen et al., 2001):

- To analyze and specify all pollution sources and their contribution to air quality
- To study the various factors, which cause the air pollution phenomenon
- To develop tools for reducing pollution by introducing alternatives for existing practices

Peak pollution episodes are a particular concern, during which ambient air concentrations are high, due to their adverse health effects for sensitive population groups such as individuals suffering from respiratory illness, children and the elderly. In Europe, the key pollutants causing the worst air quality problems are particular matter ($PM_{10}$ and $PM_{2.5}$), ozone ($O_3$) and nitrogen dioxide ($NO_2$) (Kukkonen et al., 2005). The European Union has been active in

order to foster preventive actions and regulatory measures. The Clean Air For Europe (CAFE) Directive (2008/50/EC) includes mandates for the provision of information on ambient air pollutant concentration to the public, concerning occurrences of exceedances of air quality criteria, and predictions for the next days.

### 2.4.1 Air quality forecasting

On the basis of the aforementioned issues, it is necessary to develop reliable and powerful methods for air quality forecasting (AQF), which can be used to launch preventive actions before and during the episodes. The methods can be used as part of air quality warning systems, which aim to ensure a so called early warning of urban air quality. According to International Strategy for Disaster Reduction (ISDR), United Nations, early warning can be defined as:

*"The provision of timely and effective information, through identified institutions, that allows individuals exposed to hazard to take action to avoid or reduce their risk and prepare for effective response."*

From an operational perspective, the prediction of next day's air pollution levels is usually required to launch proper actions and control strategies (Monteiro et al., 2005). In the operational setup the AQF has been previously based on numerical weather prediction (NWP), in a combination with deterministic dispersion modeling (DET) and regression-based statistical modeling. The current AQF methods are, however, limited to predict complex behavior of chemically and physically reactive air pollutants and meteorological conditions within the lowest atmospheric layer (e.g. Baklanov et al., 2002; Kukkonen et al., 2003).

In the last two decades, considerable efforts have been placed on developing advanced DDM approaches to overcome lacks of NWP/DET-based AQF. Numerous papers have been published on artificial neural networks (ANN) based AQF approaches (e.g., Nunnari et al, 1998; Kolehmainen et al., 2001; Kukkonen et al., 2003), most of them directing for the use of multi-layer perceptron (MLP) network in the prediction (e.g. Gardner and Dorling, 1998). In the accordance of the results published, the performance of ANN/MLP has been shown to be superior to that of linear modeling methods such as linear regression (e.g. Schlink et al., 2003).

In recent years, the advantages of other ANN methods such as support vector regression (SVR) for the forecasting of air quality parameters have been shown. Lu et al. (2002) and Lu and Wang (2005) have made an experimental comparison between the SVR and radial basis function (RBF) network and showed that SVR is superior to RBF in predicting respirable suspended particles (RSP), $NO_x$ and

$NO_2$. Juhos et al. (2008) evaluated the performance of SVR for predicting NO and $NO_2$ concentrations against the MLP model and found that that SVR gives more reliable forecasts, although the difference is not very substantial. Further, Juhos et al (2008) used principal component analysis (PCA) to reduce the dimensionality of the embedded input data. Chelani (2010) compared the performance of multiple linear regression (MLR), MLP and SVR in predicting $O_3$ concentrations in Delhi. The results obtained indicated the promising performance of SVR over MLP and MLR.

Moreover, wavelet-based methods have been presented. Nunnari (2003) present an approach based on wavelets for the modeling of $SO_2$ time-series. The results obtained show that there is no significant difference between the performance of wavelet-based prediction model and MLP model predictions, but that there are some advantages in using the wavelet-based method in terms of model readability. Contrary to this, the results shown by Osowski and Garanty (2007) indicate that the accuracy can be enhanced by decomposing the measured time series data into wavelet representation and predicting the lower variability wavelet coefficients of original time series using SVRs.

Promising results have been obtained also using on ensemble approaches where a number of trained ANN models share a common input and whose outputs are somehow combined to produce an overall output (Haykin, 1999). A representative example on this is presented by Siwek et al. (2010), where several ANN related modeling methods, which include MLP, SVR, RBF and Elman recurrent network, are used in parallel to forecast the daily concentrations of $PM_{10}$. In this ensemble approach, PCA is used to combine the results of individual predictors to the final neural predictor.

Despite considerable efforts with ANN-based AQF models, the evaluation of the ANN models has been largely based on "now-casting" of air quality, i.e., using the actual meteorological observations instead of NWP in the modeling (e.g. Kukkonen et al., 2003). Consequently, there is no proper understanding about the usability of a combination of NWP data and ANN methods in AQF.

Furthermore, it is often so that the building of ANN models is a long and a tedious process due to the presence of high number of potential model input variables. In this context, modern optimization methods, such as evolutionary and genetic algorithms (EA/GAs), are of particular interest, as they have not been extensively studied in the design of ANN-based AQF models. Many shortcomings are also originated from the deficiencies of air quality data. A particular issue with air quality datasets is missing data, posing many significant obstacles for the use of standard ANN models, which usually require the complete data as a condition for their use.

## 2.5 CHEMICAL RISK ASSESSMENT

Risk assessment is an important stage of environmental decision-making procedures, identifying a risk related to a concrete situation and a recognized hazard. The risk assessment process consists of the stages of:

- Hazard identification
- Exposure assessment
- Dose-response assessment
- Risk characterization

The risk assessment is associated with risk management, which is a process consisting of steps of risk classification, risk benefit analysis, risk reduction, monitoring and review. For more details on basic principles and stages of chemical risk assessment and management the reader is referred to the extensive review of Leeuwen and Vermeire (2007).

Risk assessment of chemicals is becoming more relevant after the introduction of the Registration, Evaluation and Authorization of Chemicals (REACH) regulation (2006/1907/EC). The REACH regulation requires that producers and users of chemicals have to demonstrate that their chemicals pose a low risk to the environment. Reliable risk assessment methods are therefore important in order to characterize and prevent negative impacts on health and ecosystems but also to ensure that the use of chemicals is not unnecessarily regulated. The risk assessment of chemicals is, however, a time consuming and costly process requiring often the use of laboratory (in-vivo and in-vitro) testing to identify unknown dose-responses of target chemicals.

### 2.5.1 QSARs and chemical grouping

The situation seems to change as non-testing methods (in-silico) promise considerable savings in time, money and a reduction in use of animal experiments when compared with conventional testing strategies. For example, the European Chemical Agency (ECHA) and the U.S. Environmental Protection Agency's (EPA) accept quantitative structure-activity relationship (QSAR) derived predictions for some regulatory purposes.

QSAR models are based on the similarity principle, i.e., a hypothesis that structurally similar compounds exhibit similar properties. QSAR aims to derive a quantitative model of the activity, which can be represented mathematically as follows: activity = $f$ (physicochemical properties and/or structural properties), where $f$ is a mathematical function. Biological activity (endpoint) can be expressed as the concentration of a chemical substance required to give a certain biological response, for example as lethal dose, 50% ($LD_{50}$) or lethal concentration, 50% ($LC_{50}$), of a toxin, required to kill half of a tested population

during a specified time of period. QSARs are denoted as QSPRs, when the target is a property of a chemical substance. In classical Hansch-type QSAR (Hansch et al., 1963), physicochemical parameters, steric properties or some structural features are used as descriptors. Up-to-date QSAR methods have now advanced towards more complex modeling, including the processing of 2D and 3D structure of the compounds.

QSARs relate to SARs, which are not quantitative concepts, but rather a qualitative representation of relationship based on the principle of similarity. SARs and chemical grouping are closely related methods. Chemical grouping aims to search for chemical substance groups or categories based on a structural similarity, which can be based on: common functional group, common precursor of break-down products and constant pattern in changing of potency. In REACH, the chemical grouping can be used for extracting information on more complex endpoints using the so called "read-across" approach. Annex XI of the REACH regulation defines the chemical grouping and read-across as follows:

*"Physicochemical properties, human health effects and environmental effects or environmental fate may be predicted from data for reference substance(s) within the group of by interpolation to other substances in the group (read-across approach). This avoids the need to test every substance for every endpoint."*

Despite the considerable efforts in (Q)SARs, there is still room for substantial improvement (e.g. Schultz et al., 2003). Potential pitfalls are originated from experimental data supporting the building of a model and model specification itself. Particular lacks of current (Q)SAR methods are associated with the prediction of more complex health related effects, including mutagenicity, carcinogenicity, developmental toxicity, eye and skin irritation, and skin sensitization (e.g. ECETOC, 1998; Schultz et al., 2003; Cronin and Worth, 2008). In this context, more advanced data-driven modeling approaches are of interest, as they might help to remedy the inherent limitations of current (Q)SAR methods.

## 2.6 ALS-BASED FOREST INVENTORY

Remote sensing (RS) methods are increasingly used as powerful alternatives for expensive field measurements in various applications of natural resource management. Concerning forest inventory, airborne laser-scanning (ALS) has become an important technique due to the cost-effectiveness of such methods and their accuracy relative to the current field-assessment approaches (Naesset et al., 2004). The ALS yields a three dimensional georeferenced point cloud,

which measures physical dimensions of the earth surface directly. For more details on the theory of ALS, the reader is referred e.g. to Wehr and Lohr (1999).

Most ALS-based forest-inventory methods adopt the area-based laser canopy-height-distribution approach to predict stand or plot specific forest attributes, such as mean height, basal area and stand volume, but also single-tree-based methods have been recently developed (e.g. Vauhkonen, 2010). Packalén (2010) has evaluated the methods, directing most of the efforts on the use of the non-parametric *k*-most similar neighbor method (k-MSN) (e.g. Mouer 1987; Mouer and Stage, 1995), for stand level forest inventories using ALS data and aerial photographs. Despite the relatively good prediction accuracy obtained, there is still room for improvements, especially, in respect to the extraction and the selection of appropriate ALS variables and the simultaneous prediction of species-specific forest attributes.

To improve the usability of the ALS-based forest inventory methods, advanced DDM/CI methods are of interest. In this context, ANN methods are of interest as they have been shown to be more accurate than other statistical approaches in various RS applications (e.g. Atkinson and Tatnall, 1997), but have not been extensively tested in ALS-based forest inventory.

# 3 Methods for intelligent processing of environmental data

## 3.1 HYPOTHESIS TESTING

Data analysis is a solid starting point for any kind of argument in environmental related conclusions and decision-making. Usually it is focusing on the testing of hypotheses on environmental systems using data from a controlled experiment or an observational study. According to Dowdy et al. (2004) the stages of general experimental procedure are as follows: (i) state the problem, (ii) formulate the hypothesis, (iii) design the experiments, (iv) make observations, (v) interpret the data and (vi) draw conclusions.

From another perspective, the analysis of environmental systems can be described through iterative experimental approach (Berthouex and Brown, 2002), where knowledge increases by iterating between experimental design, data collection and data analysis (Figure 3).

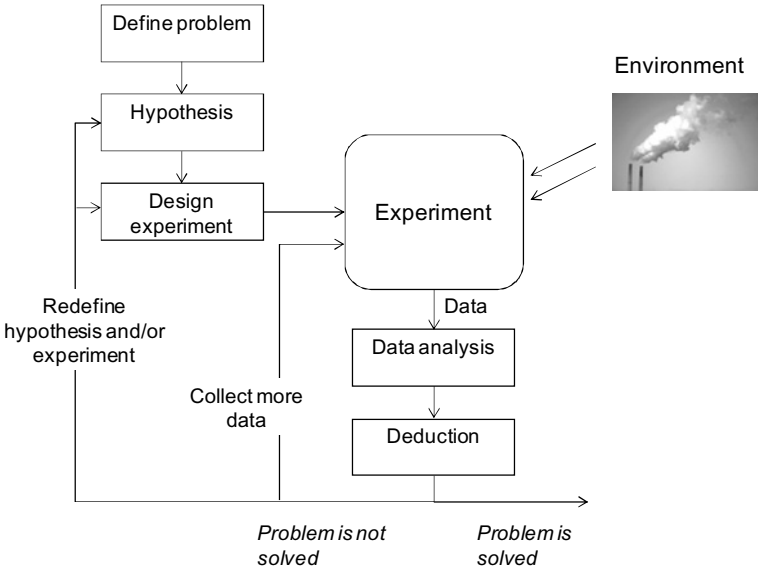


*Figure 3. The analysis of environmental systems through the elements of learning (modified from Berthouex and Brown, 2002).*

The cycle starts with the formulation of a hypothesis, which is based on a priori knowledge about the underlaying problem. The hypothesis can be represented using a mathematical model that will be used to produce predictions on the underlaying system. Iterating between data collection and data analysis provides the opportunity to enhance the model by shifting emphasis to different variables, repeating experiments and adjusting experimental settings.

However, a problem related to the analysis of environmental systems is, that it may be impossible to manipulate the independent variables to create conditions of special interest (Berthouex and Brown, 2002). A range of conditions can be observed only trough observations or field studies over a long period of time, which are not, however, necessarily collected from the same view of intention. Another problem is the replication of experimental conditions, which restrict the verification of the generated hypothesis.

## 3.2 KNOWLEDGE DISCOVERY AND DATA MINING

Environmental data are not, as previously stated, always originated from designed experiments and in many cases formulating well-defined hypotheses is difficult (Sulkava, 2008). In such conditions, the experimental procedure cannot be followed as such, but the first stage is data analysis, which could then lead to define the hypothesis, and probably also to design an experiment and collect more data (Sulkava, 2008).

The analysis of environmental systems can thus be seen as a multi-stage and iterative knowledge discovery process, in which data gathered from the underlaying system are selected, transformed and modeled in order to extract useful information (knowledge) that suggests hypothesis and conclusions, and supports decision making. Such the iterative data enrichment approach is mainly followed in this thesis.

According to Fayyad (1996), the data enrichment process, when it starts from database, can be defined as knowledge discovery in databases (KDD). In the first stage of the KDD process (Figure 4), the target data are selected for the discovery using some prior knowledge of the application domain. Next, the selected target data are undergone a preprocessing stage in which the quality of data is ensured. Usually the preprocessing covers the issues related to the handling of missing data, measurement errors and outliers. In the next stage, the data transformations and dimensionality reduction are performed in order to compress the information of the data into a smaller number of variables or a new more informative set of variables required by data mining methods. In data mining, model/patterns are extracted from data. Lastly, the model/patterns obtained in the KDD process are evaluated and interpreted.

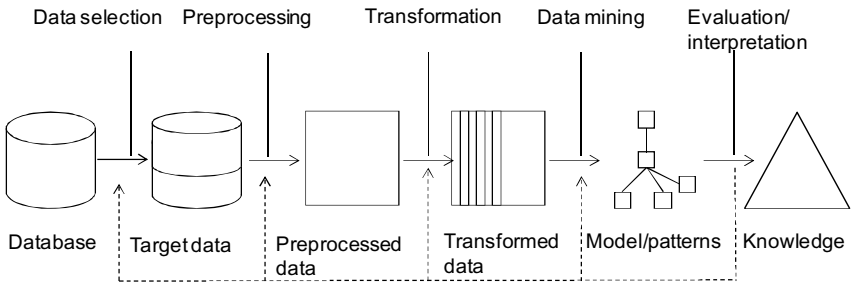


*Figure 4. The data processing chain in knowledge discovery in databases (modified from Fayyad, 1996).*

At the core of the KDD process are data mining methods. According to Fayyad (1996) the link between KDD and data mining is defined as follows:

*"KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process."*

Data mining contains a set of data analysis methods, contributed often by the methods of computational intelligence (discussed later), used to explore complex relationships and to summarize information to an understandable and useful form in data sets. Hand et al. (2001) defines the data mining as follows:

*"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."*

Data mining can be categorized into the following tasks, which correspond to different purposes of analyzing data (Hand et al., 2001):

- Exploratory data analysis
- Descriptive modeling (density estimation and cluster analysis)
- Predictive modeling (classification and regression)
- Pattern and rule discovery
- Retrieval by content

The objective of exploratory data analysis (EDA) is to explore data without a clear hypothesis of what to look for by means of simple visualization methods or more advanced data mining methods (e.g. Kolehmainen, 2004). In the descriptive modeling, intrinsic properties of the data are explored e.g. by means of density estimation and cluster analysis (Bishop, 1995; Hand et al., 2001). In cluster analysis, the data samples are portioned into subgroups according to their similarity using problem dependent proximity measures.

The aim of predictive modeling is to build a model which is able to estimate a value of one variable from values of other variables. The prediction is performed either for discrete values, when it is called classification, and for continuous function, when it is called regression (Hand et al., 2001; Han and Kamber, 2006). The rest of the data mining tasks, i.e. pattern and rule discovery and retrieval by content, are closely related for searching patterns of interest such as association rules (Hand et al., 2001).

In this thesis, the major focus was on predictive (regression) modeling. The EDA approach was considered in one of the studies as a form of qualitative approach for cluster analysis and prediction.

## 3.3 COMPUTATIONAL INTELLIGENCE

The methods of computational intelligence (CI) are increasingly used in solving complex data mining tasks in environmental sciences and engineering (e.g. Krasnopolsky and Chevallier, 2003; Solomatine, 2005; Cherkassky et al., 2006; Haupt et al., 2008). CI is an ambiguous concept, which combines the elements of learning, adaptation and evolution to create computer-based (computational) models that are, in some sense, "intelligent" (e.g. Bezdek, 1994; Fogel, 1995; Pal and Pal, 2002). According to the literature, any system that generates adaptive behavior to meet goals in a range of environments can be said to be intelligent (e.g. Bezdek, 1994; Fogel, 1995). A definition was proposed by Engelbrecht (2007):

*"Computational intelligence is the study of adaptive mechanisms to enable or facilitate intelligent behavior in complex and changing environments."*

The definition emphasizes the target, which is the complex or changing environment. Pal and Pal (2002) combine the existing definitions by requiring the following characteristics of a computational intelligent component:

- Considerable potential in solving real world problems
- Ability to learn from experience
- Capability of self-organizing
- Ability to adapt in response to dynamically changing conditions and constraints

Machine learning is a concept with respect to CI. The main goal of machine learning is to create algorithms that utilize past experience, or example data, in solving problems (Mitchell, 1997):

*"Machine learning is study of computer algorithms that improve automatically through experience."*

CI has adopted inspiration and ideas from biological mechanisms and patterns of behavior (Hanrahan, 2009). The most well-known CI methods include evolutionary and genetic algorithms, artificial neural networks and fuzzy logic. The basic principles of artificial neural networks and evolutionary and genetic algorithms studied in this thesis are shortly introduced next.

### 3.3.1 Neural networks

Artificial neural networks (ANNs) are computational models that simulate the structure and functions of biological neural networks and adopt supervised or unsupervised learning (e.g. Haykin, 1999). The ability to analyze and model complex non-linear systems makes ANNs attractive for the study of environmental systems (May et al., 2009). Basically ANN is an adaptive system consisting of a group of interconnected artificial neurons (computational units) that adapt its parameters (the connection weights) based on external or internal information that flows through the network during an iterative learning phase (training). According to Haykin (1999), a neural network can be viewed as follows:

*"A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making it available for use. It resembles the brain in two respects: (1) Knowledge is acquired by the network from its environment through a learning process, (2) Interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge."*

ANNs can be classified by the learning method (or algorithm) they are adopting. Broadly, the learning methods can be categorized into supervised learning and unsupervised (or competitive) learning methods. According to Haykin (1999), the learning can be defined in the context of ANNs as:

*"Learning is a process by which the free parameters of a neural network are adapted through a process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place."*

In supervised learning, model responses are known, and the weights of the network are adjusted so that it produces a desired input-output mapping (Figure 5). The basic aim is to infer a function, called classifier if the output is discrete, and regression function if the output is continuous, which can be used to generalize from training data to unseen external data. ANNs adopting supervised learning are particularly suitable for complex predictive modeling tasks, where the complexity of the data makes the design of a function by hand impractical.

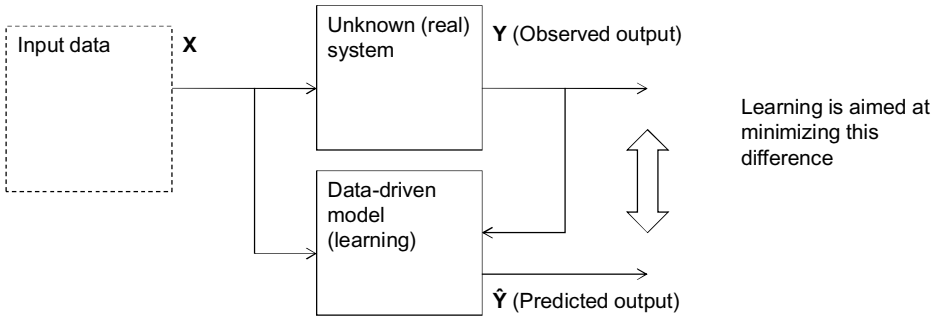


*Figure 5. The basic principle of supervised learning (modified from Mitchell, 1997).*

Some of the most well-known supervised learning ANNs are multi-layer perceptron (MLP) networks, radial basis function (RBF) networks, learning vector quantization (LVQ) and support vector regression (SVR).

It should be emphasized that ANNs adopting supervised learning include a lot of tunable parameters. It is typical that the training error tends to decrease in parallel with the increase of model complexity, and with too much fitting, called overfitting, the model captures noise in the training data, and will not generalize well (Hastie et al., 2001). As stated by Schlink et al. (2003), the presence of noise necessitates a trade-off between the accurate modeling of training data and good generalization power of the model, which is known as the bias-variance trade-off (Geman et al., 1992). The traditional way to avoid overfitting is the early-stopping, where original data are divided into the datasets of training, testing and validating. A training set is used to construct the model whereas a test set is used to control potential overfitting of the ANN model. Finally the validation set is used to evaluate the generalization ability of the model. In addition, various regularization approaches have been presented.

Conversely to supervised learning, in unsupervised learning the network aims to auto-associate information from the network inputs. Unsupervised learning methods are well-suited for approximating the probability distribution of the inputs or to discover structure in the input data (Cherrkassky and Mulier, 1998). The most well-known unsupervised ANN method is the self-organizing map (SOM), which is one of the main methods studied in this thesis.

### 3.3.2 Evolutionary and genetic algorithms

Evolutionary algorithms (EAs) comprise a class of search methods inspired by evolution and natural selection (e.g. Bäck, 1996; Fogel, 2006), and extensively used in ANN model design (e.g. Miller et al., 1998; Yao, 1999; Castillo et al.,

2002). EAs have many appealing features compared to other search and optimization algorithms, such as the ability to:

- Perform global search
- Escape local minima
- Deal with discontinuous and multi-modal functions
- Perform parallel processing (algorithm can be parallelized)

The best known EAs are the genetic algorithms (GAs) whose basic principles were first proposed by Holland (1975). GAs are iterative search heuristics mimicking natural evolution by means of selection, recombination and mutation. The theoretical background of GAs appears to be limited, but the building block hypothesis has been commonly proposed (Goldberg, 1989):

*"A genetic algorithm seeks near-optimal performance through the juxtaposition of short, low-order, high performance schemata, called building block."*

The hypothesis suggests that by decomposing the overall problem into sub-problems and solving these sub-problems separately, GA can find good solutions to the overall optimization problem.

The basic idea of GA is to create a random set (population) of bit-coded solutions (chromosomes or genotypes), which encode candidate solutions to the problem (phenotypes). Each component of chromosome represents a gene, which can be in several states, called alleles (feature values). The created population is then evolved by means of genetic search operations, namely selection, recombination and mutation, until a desired criterion is reached. The basic cycle and operations of GA are presented in Figure 6.

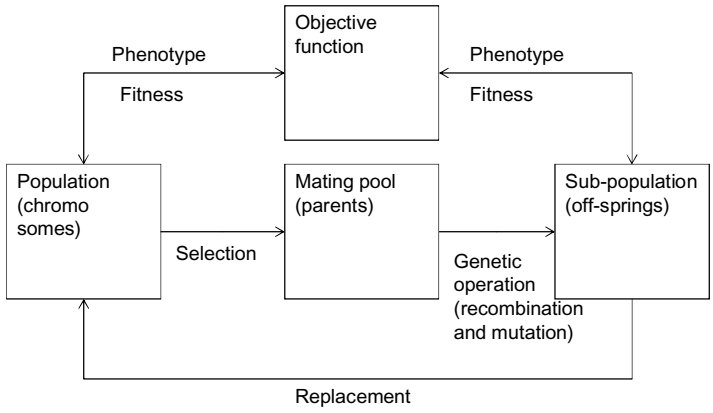


*Figure 6. The basic cycle and operations of GA (modified from Man et al. 1999).*

In the first stage of GA the population is ranked using an objective (fitness) function. According to the rank a specific ration of the population is selected in a stochastic way to reproduce a set of new solutions (off-springs) through genetic operations: recombination and mutation. Often so called elitism selection is adopted, i.e., the individual having the highest fitness selected throughout the generations. After reproducing new off-springs, follows the replacement of the old population with the new population based on the fitness. The procedure is iterated until a stopping criterion is fulfilled.

Even though EA/GAs seem to be power search algorithms in general and have been shown to often find better solutions than other search algorithms, there are some disadvantages attached to them as well, which should bear in mind. These include the selection of appropriate genetic operators and parameters, and there is no guarantee for convergence.

In addition to basic GA, numerous more sophisticated EAs have been proposed. Among the methods, there are multiple objective evolutionary algorithms (MOEAs) applied for solving complex multi-objective optimization problems. In case of multi-objective problems, no unique optimal solution can be achieved, but instead a set of trade-off (non-dominate) solutions. These solutions are known as the Pareto-optimal set (Goldberg, 1989) where no improvement in any objective is possible without sacrificing at least one of the other objectives.

Over the past decade, a number of MOEAs have been suggested, among them Fonseca and Fleming's MOGA (1995), Srinivas and Deb's NSGA (1994) and Horn et al's NPGA (1994). To attain well-distributed Patero-optimal solutions, specific Pareto ranking, sharing and goal attainment methods have been adopted. For more precise details on these methods, the reader is referred to the aforementioned references.

### 3.4 PREPROCESSING THE DATA

Preprocessing of the data is an important step in the KDD/data mining process, required to transform the data into an appropriate format required by the modeling. Typically data preprocessing deals with issues of data cleaning (e.g. handling of missing data, outliers and measurement errors), data transformations and dimensionality reduction (Han and Kamber, 2000), which are shortly discussed next.

Basically, the target data to be modeled/analyzed are often given as a data matrix, consisting of data rows and columns. The columns correspond to the measurement variables (called also attributes and features). The rows correspond to units of measurement (e.g. chemical substance or study field) or different points of time.

The format of the target data matrix is given as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad (1)$$

where $p$ is the number of variables (data columns) and $n$ is the number of samples (data lines).

Incomplete data matrices, i.e. missing data, pose most likely most significant obstacles for the adoption of standard modeling methods, which generally require complete data as a condition for their use (Gentili et al., 2003; Magnaterra et al., 2003). According to Norazian (2008) incomplete data matrices can result in three major problems which are:

- A loss of information and as a consequence, a loss of efficiency
- Complications related to data handling, computation and analysis due to irregularities in data structure and the impossibility of using standard software
- The results may be biased due to systematic differences between observed and unobserved data

Broadly there are two alternative ways to handle missing data, which are case deletion and imputation. The case deletion aims at discarding all incomplete data rows of the data matrix. However, in such an approach, significant amount of information can be lost. Contrary to the case deletion the imputation aims at replacing missing data using an imputation method. Various methods for missing data imputation have been developed (e.g. Little and Rubin, 1987; Dixon, 1979; Schafer, 1997; Schneider, 2001). However, they are often limited for dealing with inherent characteristics of environmental data.

## 3.5 DATA TRANSFORMATIONS AND DIMENSIONALITY REDUCTION

The transformation of environmental data is usually needed to transform the data into proper format required by a model. This is needed before applying a model, for the following reasons (Kolehmainen, 2004):

- The variable is cyclic, so it includes discontinuities
- The data distribution does not enable the algorithm to recover important features
- The magnitudes of the variables differ, so the variables with largest numeric values tend to dominate the modeling
- There are outliers, which suppress important features

For most standard modeling methods, cyclic variables pose a problem since they include discontinuities. Transformation of cyclic variables into continuous variables can be performed using sine and cosine functions as follows:

$$x_1 = \sin\left(\frac{2\pi x}{\omega_x}\right), \quad x_2 = \cos\left(\frac{2\pi x}{\omega_x}\right) \tag{2}$$

where $x$ is a value of the variable and $\omega_x$ describes the upper bound of the variable (e.g. for the hour of the day $\omega_x = 24$).

Many environmental variables have log-normal distributions. In some modeling circumstances, e.g. in case of standard parametric statistical methods, logarithmic transformation of variable (e.g. log10) is required to reach the assumption of normal distribution of variable.

The transformations encountered most often include data scaling, which aims to equalize the magnitudes of variables and thus preventing the variables with largest numeric values to dominate the modeling process. The standard methods include variance scaling and equalization. The variance scaling is defined as follows:

$$x' = \frac{x - x_{mean}}{x_{std}} \tag{3}$$

where $x'$ is the transformed variable, $x$ is an original value of variable, $x_{mean}$ is the mean value of variable and $x_{std}$ is the standard deviation of variable.

A particular advantage of the variance scaling is that it is not very sensitive to outliers. When the extreme values of variable are of interest, or are wanted to be weighted, more useful way could be equalization. In equalization, the variables are linearly scaled to the comparable range between 0 and 1, thus preventing any variables to dominate the modeling process:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4}$$

where $x_{min}$ and $x_{max}$ are the minimum and maximum values of variable, respectively.

Equalization can however be influenced by outliers, which are common in environmental datasets, and in such situations the variance scaling should be preferred. Further, in some situations it may be justified to use data normalization (or standardization) where the length of the data vectors is scaled to one as follows:

$$x'_{ij} = \frac{x_{ij}}{\|\mathbf{x}_i\|} \quad (5)$$

where $x_{ij}$ is an original element of the data vector and $\|\mathbf{x}_i\|$ is the length (norm) of vector $\mathbf{x}_i$.

The normalized data reflect the relative values of variable in each data vector instead of actual magnitude. This may be particularly useful when the goal is to examine rather the profiles of data vectors than the magnitudes of single variables.

It is often the case that some of variables in data are irrelevant for the modeling being carried out and dimensionality reduction is required to be carried out. Collinearity is one issue of concern, i.e. information in one of variable dublicate or overlap with information in another variable, and can lead to computational instabilities, large statistical correlations between variables and inflated standard errors (Piegorsch and Bailer, 2005). Term multicollinearity refers to a situation where two or more predictor variables in multiple regression model are linearly correlated. Another issue is the curse of dimensionality, which means that the higher the dimension the more data is needed to find accurate model parameter estimates (Sulkava, 2008).

Guyon and Elisseeff (2003) have listed the objectives of dimensionality reduction, or input variable selection, which are:

- Improving the prediction performance of the predictors
- Providing faster and more cost-effective predictors
- Providing a better understanding of the underlaying process that generated the data

When reducing the dimensionality of data, two concepts are separated. Feature selection is used to choose a subset of features, or variables, while feature extraction aims at creating a smaller set of new, more information rich, variables.

The methods can further be classified into wrapper and filter approaches (e.g. Kohavi and John, 1997; Liu and Motoda, 1998). Wrapper selection is based on iterative evaluation of models using subsets of input variables. Contrary to wrappers, filters are conversely model-free approaches where some measure is used to determine whether or not each candidate variable should be included into a set of model input variables. Since filters avoid the modeling stage, they can perform selection significantly faster than wrapper approaches (Hanharan, 2009). In addition to the filter and the wrapper approach, embedded methods have been presented. Embedded methods are usually specific to given model, performing variable selection as the part of training process. For more details on embedded methods, the reader is referred e.g. to Guyon and Elisseeff (2003).

Principal component analysis (PCA) has been commonly adopted for extracting features from environmental data (e.g. Mujunen and Minkkinen; 1996; Asikainen, 2006). PCA is a linear statistical method to project *p*-dimensional dataset **X** into a lower, *s*-dimensional space on uncorrelated variables (Jolliffe, 2002), which is often required by the standard modeling methods. The basic aim of PCA is to generate principal components (PCs) $\mathbf{TP}'$ for explaining the variance of data, so that each PC explains the maximum amount of residual variance not explained by preceding PCs. In PCA the **X**-matrix is decomposed into a sample score matrix **T**, variable loading matrix $\mathbf{P}'$ and residual matrix **E**:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \tag{6}$$

where **X** is the data matrix ($n$ x $p$), **T** is the score matrix ($n$ x $s$), **P** is the loading matrix ($s$ x $p$) and **E** is the residual matrix ($n$ x $p$), and **TP'** and **E** describes the systematic variation and the noise of data, respectively.

Sequential forward selection (SFS) is perhaps simplest algorithm for feature selection. In SFS algorithm the variables are included in progressively larger subsets until the prediction performance of the model is maximized. The corresponding selection procedure can be performed backwards, when it is called sequential backward selection (SBS).

In addition to the previous methods, sensitivity analysis (SA), called also relevance or importance analysis, can be adopted when determining most efficient variables to models. Basically, SA methods aim to evaluate model's robustness for input variables, i.e., for estimating the rate of change in a model output with respect to changes in the model inputs (Daescu, 2009). The inputs having low influence on the model output are usually considered to have low importance in the modeling and can thus be discarded.

### 3.6 EXPLORATORY DATA ANALYSIS

In exploratory data analysis (EDA) the target data are visualized and explored without a clear hypothesis by means of simple plotting of one or two variables, histograms and box plots. The aim is usually to uncover underlaying structure, extract important variables or detect outliers and anomalies from the data. However, with multidimensional environmental data, more advanced methods such as PCA, SOM and Sammon's mapping are required to reduce the dimensionality of the data (e.g. Kolehmainen, 2004).

The basic methodological basis for SOM and Sammon's mapping studied in this thesis is given next.

### 3.6.1 Self-organizing map

Self-organizing map (SOM) is one of the best known unsupervised neural learning methods (Kohonen, 2001), shown to be particularly suitable for complex environmental related data exploration tasks (e.g. Simula et al, 1999; Kolehmainen, 2004). The basic aim of SOM is to find prototype vectors that optimally represent the input data, and at the same time to achieve a low dimensional representation of the input space of the training samples usually in the two-dimensional map grid. Training of SOM results in a topological arrangement of output neurons, which are defined by their location on the map grid and by the weight vector, which has the same dimensionality as input vectors. The SOM learning is initiated by assigning random values to the weight vectors (called also prototype or reference vectors) of the network:

$$\mathbf{w}_m = (w_{m1}, w_{m2}, \ldots, w_{mp}) \tag{7}$$

where $\mathbf{w}$ is the weight vector, $m$ refers to the index of neuron and $p$ is the number of variables.

The training patterns are fed into the network one-by-one in random order, and this procedure is repeated a pre-determined number of times (epochs) for each of them. At each training step, the best matching unit (BMU) is found. The BMU is the neuron with the smallest Euclidean distance to the input vector:

$$c(\mathbf{x}_i, \mathbf{W}) = arg\min_j \left\| \mathbf{x}_i - \mathbf{w}_j \right\| \tag{8}$$

where $c$ is the index of BMU, $\mathbf{x}_i$ is the input vector and $\mathbf{W}$ includes all weight vectors. BMU and its neighboring neurons are then adjusted according to the following update rule:

$$\mathbf{w}_m(t+1) = \mathbf{w}_m(t) + h_{cm}(t)[\mathbf{x}_i - \mathbf{w}_m(t)] \tag{9}$$

where $\mathbf{w}$ is the weight vector, $m$ is the index for the neuron updated, $t$ is a counter for iterations, and $h$ is the neighborhood function. A commonly used neighborhood function is the Gaussian function:

$$h_{cm}(t) = \alpha(t) \exp\left( -\frac{\|\mathbf{r}_c - \mathbf{r}_m\|^2}{2\sigma^2(\mathrm{t})} \right) \tag{10}$$

where $\alpha(t)$ is a learning rate factor, $\mathbf{r}_c$ and $\mathbf{r}_m$ are the location vectors in the map grid for the corresponding neurons and $\sigma(t)$ defines the width of the kernel.

To summarize, the SOM algorithm proceed as follows:

(i) Find the BMU for input vector according to the minimum Euclidean distance (Eq. 8)
(ii) Move the weight vector of the BMU towards that input vector, using the update rule (Eq. 9)
(iii) Move the weight vectors of neighboring neurons, according to the neighborhood function (Eq. 10), towards that input vector, using the update rule (Eq. 9)
(iv) Repeat steps (i)–(iii) for the next input vector until all input vectors have been used
(v) Repeat steps (i)–(iv) until convergence
(vi) Find the final BMU for each input vector according to the Euclidean distance

In principle, SOM is particularly suitable for different data exploration tasks, but it can be used also for regression modeling. Regression can be accomplished by searching the BMU for the vector with unknown components using the known vector components. In this thesis, the SOM was used for regression in two of the studies (**Papers I** and **V**). The output of SOM was based on the mean value of target variables of the BMU for the selected predictor variables, following the principle adopted by Kolehmainen et al. (2001).

### 3.6.2 Sammon's mapping

Sammon's mapping (Sammon, 1969) is a non-linear mapping tool that belongs to the so-called metric multidimensional scaling methods (Kohonen, 2001). Sammon's mapping is a useful tool for the preliminary analysis and visualization of class distributions and the degree of their overlap. Sammon's mapping aims at representing the points of $p$-dimensional data onto a subspace of two dimensions, preserving, however, the inter-pattern distances as far as possible.

Basically Sammon's mapping is based on the minimization of the following cost function:

$$E = \frac{1}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{\left(d_{ij} - d'_{ij}\right)^2}{d_{ij}} \qquad (11)$$

where $n$ is the number of data points, $d_{ij}$ is the Euclidean distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the original space, and $d'_{ij}$ is the Euclidean distance between the corresponding points $\mathbf{x}'_i$ and $\mathbf{x}'_j$ in the lower dimensional target space.

The minimization is usually based on the steepest descent (Kohonen, 2001), updating the positions in the target space as follows:

$$x'_{ip}(t+1) = x'_{ip}(t) - \alpha \frac{\partial E(t)/\partial x'_{ip}(t)}{\left|\partial^2 E(t)/\partial x'_{ip}(t)^2\right|} \quad (12)$$

where $x'_{ip}$ is the $p$th coordinate of the position of the point in target space, $t$ is the counter for iterations and the factor $\alpha$ is experimental factor, influencing on the convergence of the algorithm.

Contrary to PCA, Sammon's mapping is capable of maintaining the non-linear properties of data. In addition, the benefit of Sammon's mapping is the ability to compress highly multidimensional and collinear information into very low dimension. Unfortunately, the numerical calculation of Sammon's mapping is time-consuming, which significantly restricts its usage for large datasets. In such data conditions, the combination of SOM and Sammon's mapping has been shown to be an efficient alternative (Kolehmainen, 2004).

### 3.7 PREDICTIVE MODELING

According to Hand et al. (2001) predictive modeling can be seen as a data mining task, which aims to build a model, which is able to estimate a value of one variable from values of other variables. The basic goal is to find a model, which fits the training data and produces maximal accuracy (low bias) and precision (low variation) with the external validation data (Berthouex and Brown, 2002). In this context, the selection of the appropriate model structure is of particular importance to obtain good generalization. With an insufficient amount of model parameters it is not possible to get a good fit to the data, whereas if too many parameters are used, the model fits the training data but has poor external performance (generalization) with external data.

Following Åström and Wittenmark (1990), the main components of the model building include: (i) selection of the model structure, (ii) parameter estimation, and (iii) model validation. Selection of the model structure aims at determining the model input-ouput signals and the internal components of the model, appropriate to the problem studied. In parameter estimation, the values of the unknown parameters of a model structure are estimated using parameter estimation methods. The selection of the parameter estimation method depends on the structure of the model, as well as the properties of the data. Lastly, in model validation, the goodness of the model is assessed, usually in respect to accuracy and generalization ability.

Next, the regression modeling methods studied or referred to in this thesis are briefly described and discussed.

### 3.7.1 Conventional regression methods

The most well-known regression method is linear regression (LR). Basically, LR aims at predicting variables as a function of a set of some observable predictor variables (Piegorsch and Bailer, 2005). The conventional multiple linear regression (MLR) model can be expressed in matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{13}$$

where $\mathbf{y}$ includes the values of the dependent variable (or response variable), $\mathbf{X}$ includes the values of independent variables (or regressors, or explatory variables, or predictor variables, or input variables), $\boldsymbol{\beta}$ are the regression coefficients (parameters) and $\mathbf{e}$ includes the residual errors (e.g. due to noise).

The parameters of (M)LR models are often estimated using the least squares (LS) approach, which is based on minimizing the sums of squared residuals. The LS estimates of the regression coefficients can be solved using the following matrix algebra called pseudo-inverse (Berthouex and Brown, 2002):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{14}$$

For non-linear equations, an algebraic solution cannot be found, and parameter estimation is usually carried out using iterative methods such as Gauss-Newton (Berthouex and Brown, 2002).

With highly dimensional and collinear data, PCA-based regression methods are extensively adopted (e.g. Mujunen and Minkkinen, 1996). Commonly adopted methods include principal component regression (PCR) and partial least squares (PLS) (Geladi and Kowalski, 1986; Esbensen, 2002). In PCR, orthogonal PCs are used straightforwardly as the independent variables instead of the original variables, which avoid the use of collinear data in the modeling. In PLS, both input and output matrices are decomposed into score, loading and residual matrices, and PCs are generated as linear combinations of original variables in such a way that they provide maximum correlation with the dependent variable.

Environmental processes are usually described through autocorrelated time-series. The existing correlation can be used to predict future behavior of a variable on the basis of past records of variable and/or other external (exogenous) variables. The classical methods to analyze and predict such series include autoregressive (AR) and moving average (MA) models. In principle, AR is an application of linear regression, in which a linear model is formed using

previous values of the same variables. The method is called ARX, if the model includes one or more external variables (exogenous). For more details on time-series models (such as AR/ARX) and their applications to dynamic systems the reader is referred e.g. to Ljung (1999).

It is often inadequate to adopt linear regression models, due to non-linear characteristics of environmental systems. To overcome this problem, various non-linear regression methods have been developed (e.g. Piegorsch and Bailer, 2005). These methods include e.g. piecewise regression model, also known as a segmented regression model, based on segmentation of the modeling along the range of the predictor variable, exponential regression models based on expected exponential functional relationships, growth curves, rational polynomials and multiple nonlinear regression (Piegorsch and Bailer, 2005). Other non-linear regression modeling methods include artificial neural networks (ANNs), such as multi-layer perceptron (MLP) and support vector regression (SVR), studied in this thesis.

As alternative to the previous parametric regression methods, non-parametric regression methods can be used. In non-parametric regression no underlying parametric model is assumed, but only a large amount of the data. It relies on assumption that the value of an unknown sample can be predicted using the values of its nearest neighbors. In most of the cases Euclidean distance metric is adopted to search nearest observations. Nearest-neighbor regression (NN) is probably the simplest and computationally easiest non-parametric regression method. In addition, there are other more sophisticated approaches, such as the most similar neighbor (MSN) method (Mouer, 1987).

### 3.7.2 Multi-layer perceptron

The multi-layer perceptron (MLP) is the most commonly used feed-forward neural network, having numerous applications to prediction, function approximation and classification in environmental sciences (e.g. Gardner and Dorling, 1998).

The MLP network consists of processing elements, called neurons or nodes, and connections (Haykin, 1999). The processing elements are arranged as layers, the input layer, hidden layer(s) and output layer. An input layer distributes input signals to the hidden layer. Each unit in the hidden layer sums its input, processes it with a transfer function (called also activation function), and distributes the result to the output layer, or in case of several hidden layers, to the next hidden layer. The units in the output layer compute their output in a similar way. Usually the sigmoidal transfer function is used in the hidden layer and linear transfer function in the output layer when modeling a continuous function.

The basic principle of a neuron model is illustrated in Figure 7, where the output signal of a single neuron is expressed as:

$$y_j = f\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right) \quad (15)$$

where $f$ denotes the transfer function, $j$ is the index of the neuron, $n$ is the number of neurons in input layer, $x_i$ is the input from $i$th input neuron, $w_{ij}$ is the weight between $i$th input neuron and $j$th hidden neuron and $b_j$ is the bias of the neuron.

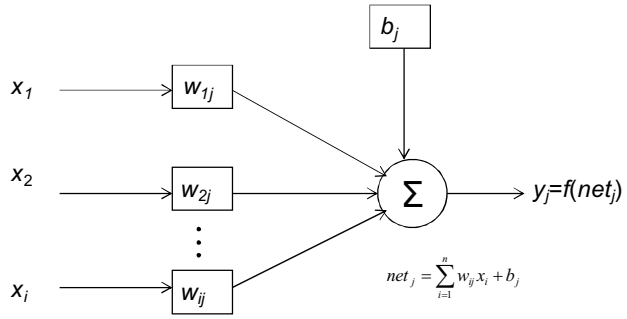


*Figure 7. Basic principle of a neuron model.*

It has been shown that the MLP network is an universal approximator, i.e. capable of approximating any measurable function to any desired degree of accuracy (Hornik et al., 1989). Thus, specific attention is required to be placed on the selection of appropriate model structure, in order to prevent a risk of overfitting and to achieve sufficient generalization. Usually one hidden layer is shown to be sufficient approximation in regression problems.

Training of the MLP network is performed using the back-propagation (BP) algorithm (e.g. Bishop, 1995; Haykin, 1999) by adjusting iteratively the weights of the network to minimize the network error function, i.e., a sum of squared errors calculated between actual and desired outputs for all data input rows:

$$E(\mathbf{w}) = \sum_{i=1}^{n} e_i^2(\mathbf{w}) \quad (16)$$

where $n$ is the number of data input rows, $\mathbf{w}$ contains the weights and biases of the network, and $e_i(\mathbf{w})$ contains the error of the network for input row $i$.

In the first phase (called forward pass), the signals of the network are computed starting from the input layer and resulting in the output layer. Next the difference between the computed and known (i.e. measured data) output is calculated. This error signal is then propagated backwards in the network in the

second phase by calculating the local gradients of the neurons and adjusting each weight value according to the local gradient and the current signal value.

At general level, i.e. not paying attention to the details of the implementation, it is possible to describe the learning using the well-known formula of gradient descent as follows:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha(t)\mathbf{g}(t) \tag{17}$$

where $\alpha(t)$ is a learning rate factor, $\mathbf{w}(t)$ is a vector of current weights, $\mathbf{g}(t)$ is the current gradient for weights and $t$ is a counter for iterations.

Major problems associated with the basic BP algorithm are, however, slowness in learning, local minima and poor generalization. To overcome the previous drawbacks several enhancements have been proposed, including the methods of numerical optimization (e.g. Haykin, 1999). Some of the methods are based on the computation of Hessian matrix, which contains the second derivates of the network errors with respect to the weights and biases. Unfortunately, the computation of the Hessian matrix is computational demanding for feed-forward networks, and thus the methods, which do not require the computation of second derivates, have been developed.

Perhaps, the most efficient algorithm for medium size networks is the Levenberg-Marquardt (LM) algorithm (Hagan and Menhaj, 1994), which is also used in comprising works of this thesis. The LM algorithm was designed to approach a second-order training speed without requiring computation of the Hessian matrix. When the error function has the form of a sum of squares, then the Hessian matrix can be approximated:

$$\mathbf{H} = \mathbf{J}^T\mathbf{J} \tag{18}$$

where $\mathbf{J}$ is the Jacobian matrix containing first derivates of the network errors with respect to the weights and biases:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial e_1}{\partial w_1} & \cdots & \frac{\partial e_1}{\partial w_{\mathrm{N}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial e_n}{\partial w_1} & \cdots & \frac{\partial e_n}{\partial w_{\mathrm{N}}} \end{bmatrix} \tag{19}$$

where the N is the number of weights.

Next the gradient of the error function can be computed:

$$\mathbf{g} = \mathbf{J}^T\mathbf{e} \tag{20}$$

where $\mathbf{e}$ is a vector of networks errors $[e_1, e_2, \ldots, e_n]^T$. The LM algorithm uses the approximation of $\mathbf{H}$ in the following Newton-like update:

$$\mathbf{x}(t+1) = \mathbf{x}(t) - [\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}]^{-1}\mathbf{J}^T\mathbf{e}(t) \tag{21}$$

where $\mu$ is a scalar and $\mathbf{I}$ is the identity matrix.

Following Hagan and Menhaj (1994), the parameter $\mu$ is multiplied by some factor $\beta$ when a step would increase the network performance index and divided by it when a step decreases the performance index. Suitable initial values could be $\mu$ = 0.01 and $\beta$ = 10, as suggested by Hagan and Menhaj (1994).

### 3.7.3 Support vector regression

Support vector regression (SVR) is a modern regression method closely related to feed-forward neural networks (Haykin, 1999; Vapnik, 1995; Burges, 1998). In recent years, SVR has been increasingly used in the field of environmental modeling, e.g., Yu and Liong (2007) in forecasting hydrologic time-series, Mileva-Boshkoska and Stankovski (2007) in predicting ozone concentrations, Lu and Wang (2005) in $NO_x$ and $NO_2$ prediction, Lu et al. (2002) in predicting respirable suspended particles (RSP), and Canu and Rakotomamonjy (2001) in predicting $O_3$ concentrations.

SVR adopts the structure minimization principle, which has been shown to be superior to the traditional empirical risk minimization employed by conventional neural networks. The theory of SVR originates from support vector machines (SVMs), which are developed for classification task. The most commonly used implementation is $\varepsilon$-SVR (e.g. Drucker et al. 1997; Vapnik, 1995), which is basically an extension of the linear regression model, which aims to find the following function:

$$y = \mathbf{w}^T\mathbf{x} + b \tag{22}$$

The learning task is transformed to the quadratic optimization problem based on the minimization of the so-called Vapnik's $\varepsilon$-insensitive loss function (for more details see e.g. Smola and Schölkopf, 1998 and Vapnik, 1995). The optimization problem can be formulated into the following form:

$$min\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (23)$$

Subject to:

$$y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \varepsilon + \xi_i$$

$$\mathbf{w}^T\mathbf{x}_i + b - y_i \leq \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

where $n$ denotes the number of samples, $C$ is a positive constant that defines the degree of penalized loss when a training error occurs, i.e., trade-off between the training error and the model flatness, $\varepsilon$ is the radius of the insensitive zone and $\xi$ are slack variables to measure the deviation of training samples outside $\varepsilon$– intensive zone.

After solving the optimization problem, the following function can be used to estimate new data points:

$$y = \sum_{i=1}^{n}(\alpha_i + \alpha_i^*)\langle \mathbf{x}_i, \mathbf{x}\rangle + b \quad (24)$$

where $\alpha_i$, $\alpha_i^*$ are Lagrange multipliers (for the support vectors $\alpha_i$ or $\alpha_i^*$ are non zero) and $\langle .,. \rangle$ denotes the dot product.

This is the so-called support vector expansion in which **w** can be described as a linear combination of the training patterns. In a sense, the complexity of a function is independent of the dimensionality of input space but depends only on the number of support vectors, which are a small subset of training data extracted by the algorithm.

The non-linear property of SVR is achieved by mapping of the input vector **x** into the higher dimensional feature space, using a non-linear mapping function $\varphi(\mathbf{x})$:

$$y = \sum_{i=1}^{n}(\alpha_i + \alpha_i^*)\langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x})\rangle + b \quad (25)$$

To obtain sufficient efficiency, the mapping can be performed using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ to yield the inner products in the feature space rather than calculating $\varphi(\mathbf{x})$ explicitly:

$$y = \sum_{i=1}^{n}(\alpha_i + \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b \quad (26)$$

The radial basis kernel function is commonly used, which is given as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right) \quad (27)$$

where $\sigma$ is the spread parameter.

Most important is the choice of control parameters, i.e., coefficients $\varepsilon$ and $C$ (Cherkassy and Ma, 2004; Chen and Wang, 2007; Osowski and Garanty, 2007). The regularization constant $C$ controls the smoothness of the approximation function, i.e. a greater $C$ value indicates less regularization and a more complex model. The constant $\varepsilon$ determinates the margin within the error is neglected, dominating the number of support vectors by governing the accuracy of the approximation function. For normalized input signals the value of $\varepsilon$ is usually adjusted in the range ($10^{-3}$–$10^{-2}$) and C is much bigger than 1 (Osowski and Garanty, 2007).

### 3.8 MODEL VALIDATION

The evaluation of the model's generalization ability is an important step of the predictive modeling. In principle, the model validation is performed using an external, independent validation data, which has not been incorporated to the building stage of the model. In addition, the internal performance of the model (goodness-of-fit) is usually relevant to examine.

Commonly used validation methods include hold-out, *k*-fold cross-validation (leave many out, LMO) and leave one out (LOO) (e.g. Snee, 1997; Michaelsen, 1987). Moreover, the method of bootstrapping is used for re-sampling of the validation set to produce the distribution of re-sampled validation indices (Efron and Tibshirani, 1993). In the holdout scheme, the data are randomly divided into training and validation sets. The validation set is used to test the performance of a model built on training data. However, such a method can underestimate the prediction power of a model due to insufficient sampling, i.e., how the data is divided into training and validation set. Opposite to the holdout, the LMO divides the data into several subsets which are in turn used as a validation set and the rest of the subsets as the training set. The LMO method enhances the statistical reliability of the performance estimate compared to the hold-out method. The basic idea of the LOO is similar to the LMO but it tries to maximize the amount of the training data by testing a model for each data row.

The holdout and LMO methods are commonly used in case of large environmental time-series datasets. The benefit of these methods is the computational efficiency compared to the LOO and bootstrapping. The LOO

and the bootstrapping are suitable methods for validating the models limited to relatively small number of data rows. The selection of a feasible method should be, however, always made case by case, and it is recommended to use several methods simultaneously to achieve more extensive understanding about the performance.

Several statistical measures have been presented for the measuring of performance of a predictive regression model (e.g. Willmott, 1981; Willmott et al., 1985). In principle, the validation statistics are based on the calculation of validation error $e$, i.e. the difference of the observed data point $y_i$ and predicted data point $\hat{y}_i$ for data lines $i$, …, $n$ in the validation set:

$$e_i = y_i - \hat{y}_i \tag{28}$$

Most likely the most common statistical measure is the coefficient of determination ($R^2$), which indicates how much of the observed variance is accounted for by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{29}$$

where $\bar{y}$ is the observed mean of variable.

However, there are defects with $R^2$ when using it for evaluating and inter-comparing models. For instance, in certain situations the magnitude of $R^2$ is not consistently related to the accuracy of the prediction (e.g. Fox, 1981; Willmott, 1981). This is the case e.g. when the estimates $\hat{\mathbf{y}}$ correlate well with the measurements $\mathbf{y}$, but a systematic offset is observed.

Fox (1981) recommended for calculating the mean absolute error (MAE), the mean bias error (MBE) and the root mean square error (RMSE). The MAE is calculated simply as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{30}$$

where $n$ is the number of observations in the validation set.

To calculate RMSE, sums of squares of errors (SSE) or the predicted residual sum of squares (PRESS, Weisberg, 1985) is determined first as follows:

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 \tag{31}$$

SSE can be further transformed to the mean squared error (MSE), which is the average squared error for the validation set divided by the number of observations:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n} \tag{32}$$

From MSE RMSE can be calculated:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{33}$$

The advantage of RMSE over MSE is that it is in the original units of the estimated variable. RMSE can be divided into its systematic ($\mathrm{RMSE_s}$) and unsystematic ($\mathrm{RMSE_u}$) components using a least-squares estimate of the predicted data point. Then to describe how much the model underestimates or overestimates the values (the bias), the mean bias error (MBE) can be determined as follows:

$$\mathrm{MBE} = \frac{\sum_{i=1}^{n} e_i}{n} \tag{34}$$

To get a relative and dimensionless measure of the accuracy the index of agreement (IA, called also *d*) can be calculated as follows (Willmott, 1981):

$$\mathrm{IA} = 1 - \left( \frac{\mathrm{SSE}}{\sum_{i=1}^{n} (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \right) \tag{35}$$

In general, IA is an appropriate and well-understandable operational measure limited to the range of 0–1, i.e., if it is not good then it is unlikely that the model can be used in practice (Kolehmainen, 2004).

When the aim is to model rare environmental events, conventional validation statistics cannot solely guarantee the performance of the model (Cherkassky et al., 2006). Methods for evaluating models in such critical situations (e.g. urban air pollution episodes; Schlink et al., 2003) are the fraction of false predictions (FA), the fraction of correct predictions (TA) and the success index (SI):

$$\mathrm{SI} = \mathrm{TPR} - \mathrm{FPR} \tag{36}$$

where TPR is the true positive rate representing the sensitivity of the model (the fraction of correct predictions) and FPR is the false positive rate, representing the specificity of the model. SI is limited to the range of -1–1 and for a perfect model SI = 1.

The significance of the model predictions can be evaluated using various statistical tests. A commonly used test is the F-test, which is used to assess the overall significance of the regression model. The F-value is the ratio between explained model variance (systematic part) and unexplained model variance (random part):

$$F = R^2(n - p - 1)/(p(1 - R^2)) \quad (37)$$

The estimation of the standard errors for the estimated parameter can be performed using the method of bootstrapping. The bootstrapping is a non-parametric approach utilizing the re-sampling of the validation set with replacement and calculating the indicators for each set separately to produce the distribution of re-sampled validation indices (Efron and Tibshirani, 1993). The standard error for the estimated parameter is thus the standard deviation of the re-sampled indices.

# 4 Case studies

## 4.1 AIMS OF THE PRESENT STUDY

The objective of this thesis was to evaluate the usability of modern computational methods and related DDM approaches for solving complex predictive modeling tasks associated with environmental management decision-making. The thesis was carried out through the representative case studies, which fall into the following fields of modeling: (i) air quality modeling and forecasting (**Papers I–III**), (ii) QSARs and chemical grouping (**Paper IV**) and (iii) remote sensing (**Paper V**). The specific research objectives studied in each application domain are as follows:

- To evaluate and compare the performance and suitability of SOM and MLP and other computational methods for recovering missing data in air quality datasets (**Paper I**)
- To evaluate the performance of MLP-based modeling schemes for the forecasting of hourly urban air pollutant concentrations (**Papers II–III**)
- To evaluate the usability of a chemical grouping approach based on Sammon's mapping and regression-based QSAR models for predicting physicochemical and biological properties of a set of target chemicals (**Paper IV**)
- To evaluate the accuracy of MLP, SVR and SOM for the prediction of species-specific stem volumes using ALS and other remote sensing data (**Paper V**)
- To test the usability of GA-based schemes for selecting appropriate model structure and input variables in the forecasting of hourly air quality (**Papers II–III**) and ALS-based forest inventory (**Paper V**)

In each application domain, the data-driven modeling approaches were designed and evaluated through an examination of the external validity using a comparison of model output with the experimental data.

## 4.2 EXPERIMENTAL DATA

Experimental data were derived in each application domain, varying from continuous monitoring data to more static field or laboratory experiments. The characteristics of the experimental data posed many requirements for the computational methods and approaches used. The experimental data and its properties are briefly summarized in Table 1.

*Table 1: The properties of experimental data sets used in this thesis.*

| **Paper** | **I** | **II–III** | **IV** | **V** |
|---|---|---|---|---|
| Data source | Air quality monitoring stations and meteorological stations | Air quality monitoring stations and meteorological stations; NWP | Laboratory in-vitro and in-vivo experiments | Remote sensing and field experiments |
| Number of data lines | ~8758 (1 year) | 25 000 – 35 000 (3-4 years) | 32 | 463 |
| Number of data columns (variables) | Less than 15 | Between 30 and 100 | Up to 1191 | 44 multiplied by transformations |
| Temporal resolution | 1 hour | 1 hour; 6 hours (NWP data) | (no temporal data) | (no temporal data) |
| Spatial dimension | ~30 x 30 km | ~30 x 30 km | (no spatial data) | < 10 x 10 km |
| Data quality | 0 –15% missing data | 0 –15% missing | Erroneous values (constants and zeros) | (no severe data quality problems) |

Air quality time-series datasets were investigated in **Papers I–III**. The data sets consisted of concentrations of major airborne pollutants monitored in urban traffic and urban background stations, all in hourly averaged time-scale, and processed according to the existing quality assurance/quality control (QA/QC) procedures. The concentration data were supplemented with the basic meteorological variables.

In **Paper I** the experimental data consisted of airborne concentration data and meteorological observation data monitored in Helsinki, Finland and Belfast, Northern Ireland, during the year 1998 (Sect. 2.1, **Paper I**). In **Papers II** and **III** the concentration and meteorological data were monitored in the Helsinki metropolitan area during the periods of 1996–1999 (Sect. 2.1, **Paper II**) and 2000–2003 (Table I, **Paper III**), respectively, and supplemented with additional parameters such as the Monin-Obukhow length and the mixing height, estimated using the meteorological pre-processing model MMP-FMI (e.g. Karppinen et al., 2000). In **Paper III** the meteorological input data were supplemented with the NWP data produced by the HIRLAM limited area weather forecasting model (Eerola, 2002) (Sect. 2.1.2, **Paper III**). The HIRLAM grid point nearest to the selected air quality monitoring stations was selected. For this point, all the forecasts (listed in Table I, **Paper III**) from the model surface levels made within 00, 06, 12 and 18 Coordinated Universal Time (UTC) were employed.

In **Paper IV** the chemical descriptor data associated with a group of chemical substances were examined. The chemical substances are presented in Table 3,

**Paper IV**. The data set consisted of molecular descriptors (topological, geometrical, connectivity indices, etc.) calculated from chemical molecular structures using the DRAGON package (Talete, http://www.talete.mi.it), comprising altogether a total of 1191variables. Furthermore, some other variables were calculated using standard packages (e.g., ALOGPS 2.1, HYPERCHEM) and methods such as EVA (Turner and Willett, 2000). The data produced by the molecular modeling were linked with the corresponding physicochemical (Table 4, **Paper IV**), fate and toxicity variables (Table 5, **Paper IV**) derived by means of the standard EPA/ECOSAR models (http://www.epa.gov) or simple linear regression models (e.g. Papa et al., 2005; Gramatica et al., 2007).

In **Paper V,** the remote sensing (RS) data supplemented with the experimental field measurements (II A, **Paper V**) were collected from 463 sample plots in 67 randomly chosen stands, in the Matamansalo test area in Eastern Finland. The RS data consisted of the feature data calculated from ALS data and digitized and corrected aerial photographs (II B, **Paper V**). Species-specific volume estimates were calculated for each plot as a function of measured diameter at breast height (DBH) and calculated tree height (Veltheim, 1987) using the species-specific models presented by Laasasenaho (1982).

## 4.3 COMPUTATIONAL APPROACH

The general computational approach adopted follows mainly the stages of the previously presented KDD process (Fayyad, 1996), considering the data as a static resource of the information for the modeling. The main stages of the data processing and modeling carried out can be condensed into the following chain:

- Data collection and selection
- Data preprocessing (cleaning, transformations and dimensionality reduction)
- Model parameter and model structure selection
- Model validation
- Model interpretation

In the first stage, the experimental data gathered from each application domain was preprocessed and transformed to appropriate format for the modeling. The actual (predictive) modeling step was then performed using various computational modeling methods. In the modeling stage, the selection of appropriate model input variables and structure was performed in order to enhance the predictive capability of a model. After which, the external validity of the modeling was evaluated against experimental data using the standard validation methods and performance measures. Lastly, the interpretation of the

resulting models was performed by utilizing the problem-specific expert knowledge in each application domain.

### 4.3.1 Data preprocessing

In **Papers I–III**, the imputation of missing data was performed in order to ensure the use of the ANN-modeling methods. This was followed by the transformation of discontinuous timing variables (such as day of week and hour) and wind direction into continuous series of sine and cosine components. In **Paper V**, various transformations of original features were calculated for the basis of input variable selection. The data scaling was performed next using the standard variance scaling, as suggested to be robust for potential outliers. In **Paper IV**, Sammon's mapping was used to reduce the dimensionality of the data in order to facilitate the interpretation of the modeling results.

### 4.3.2 Modeling methods

The selection of modeling methods was made based on the previous studies and knowledge, which in general suggest the potential of modern computational data-driven methods in solving complex and ill-defined environmental modeling tasks (e.g. Kolehmainen et al., 2001; Canu and Rakotomamonjy, 2001; Lu et al., 2002; Kolehmainen, 2004; Lu and Wang, 2005). The main emphasis has been on the up-to-date ANN models, namely: MLP, SVR and SOM. The methods were used in parallel or in a combination with the other statistical modeling methods in order to create problem tailored approaches.

In **Paper I**, the MLP and SOM were benchmarked for filling-in the missing air quality data. The methods were benchmarked with other statistical imputation methods including linear interpolation (LIN), multiple linear regression (MLR) and multivariate nearest neighbor (NN) regression (Dixon, 1979). In **Papers II** and **III**, MLP-based air quality forecasting (AQF) schemes were studied. Previously, the ability of the MLP network for the forecasting of concentrations of a range of pollutants has been shown by many studies (e.g. Gardner and Dorling, 1998; Kolehmainen et al., 2001; Kukkonen, et al., 2003). In **Paper IV**, Sammon's mapping was adopted for the discovery of chemical substance groups from high-dimensional chemical descriptor data. Previously, the benefits of Sammon's mapping for solving complex data exploration tasks have been shown, for instance for the analysis of urban air quality and fermentation process data (Kolehmainen, 2004). Lastly, in **Paper V**, MLP, SOM and SVR were applied for predicting species-specific stem volumes using ALS data and airborne photographs, benchmarking the methods in respect to the non-parametric k-MSN method applied previously in the domain (e.g. Packalén and Maltamo, 2007).

### 4.3.3 Variable and parameter selection

The selection of appropriate ANN input variables and structure was of particular concern since a too complex network can lead to over-fitting and poor generalization power. The major emphasis was on testing GA-based selection schemes, which have been shown to merit many appealing benefits in ANN design, such as the ability to deal with large feature and model architecture space, non differentiable and noisy error function and multimodality (e.g. Miller et al., 1998; Yao, 1999; Castillo et al., 2002). For more throughout information about the details of the developed GA-based schemes, the reader is referred to the original research papers (**Papers II** and **III)**.

In case of the MLP network (**Papers II**, **III** and **V**) the model structure selection was formulated using the direct or in-direct encoding. In the direct encoding each phenotypic feature (e.g. network input) is encoded by using exactly one genotypic code, whereas in the indirect encoding only some characteristics of the model are encoded (e.g. Yao, 1999). The selection of model input variables was implemented, using the direct encoding by representing the model input variables as a bit string, where 0 refers to the absence of an input variable and 1 refers to the presence of an input variable in the model. Contrary to this, the indirect encoding was used for selecting the high-level architecture of the MLP network, by encoding the number of hidden layers and nodes to the bit string (see Figure 1, **Paper II**). The benefit of the indirect encoding is its better scalability, but on the other hand small changes in the representation (e.g. number of hidden layers) might lead to significant changes in fitness (Castillo et al., 2002).

The evaluation of generalization ability for different MLP network structures was performed using the conventional hold-out validation, in which the performance of the trained networks was assessed using IA, calculated based on external validation set (**Paper II**). The conventional early-stopping strategy was used to control over-fitting by comparing the error of the training set to the corresponding error of the validation set during the training (Sarle, 1995). However, in practice, the evolving of the MLP network is computationally tedious, due to the iterative training phase of MLP. To overcome this problem, the actual training of MLP was replaced by a sensitivity analysis (SA) in the MLP design stage (**Paper III**). In this scheme, the sensitivity of an input was estimated by replacing an input variable in the test set by its average computed on the training set and calculating the effect of this elimination on the output of the MLP network.

In case of the SVR model, the complexity of model structure was controlled straightforwardly using hyper parameters $\varepsilon$ and C, which control noise-sensitivity and flatness of the model, respectively. In case of SOM, major efforts focused on the selection of appropriate SOM size (**Papers I** and **V**). This was

performed mainly through experimental testing of different alternatives for SOM size in respect to the quantization error or the prediction error of the SOM model.

### 4.3.4 Model validation

The validation of model's external prediction performance is an essential stage of the model building process. In this thesis, the prediction performance was evaluated using CV/LOO statistics based on the standard statistical performance measures and their bootstrapped confidence levels. The selection of validation methods was made based on the previous studies in the application domains studied.

In air quality modeling, the cross-validation between the years has been largely used to test the ability of air quality models to predict values outside the training years (e.g. Kukkonen et al., 2003). In QSAR modeling and ALS-based forest inventory the LOO statistics have been commonly calculated to describe the prediction ability of the models (e.g. Packalén and Maltamo, 2007; Gramatica et al., 2007). In case of QSAR modeling also internal performance rates, reflecting goodness-of-fit, were reported (**Paper V**).

The standard RMSE and $R^2$ were calculated throughout the experiments, supplemented with the dimensionless IA. To describe how much the model underestimates or estimates the response variable, the bias was calculated. SI was used to determine the success in the prediction of critical events (**Papers II** and **III**). Further, to assess the statistical significance of the models, the standard *t*-test or F-test was employed in some conditions (**Papers IV** and **V**).

## 4.4 IMPLEMENTATION OF THE MODELING SCHEMES

The data processing and modeling approaches covered in this thesis were implemented mainly using in-house Matlab functions and scripts based on available Matlab toolboxes. Based on that programming work the Matlab toolbox (called MTools) was designed and implemented (currently not publically available). The contribution of MTools was that of facilitating and speeding up the required model design process, as the whole data processing chain was not needed to be coded always from scratch throughout the experiments.

In principle, MTools library is composed of the user defined run script, the main function and low-level data processing and modeling functions (Figure 8). MTools main function covers all the relevant stages of data processing required to build up and test predictive data-driven models for the target data. Actual

modeling methods are based on the implementation of the existing Matlab toolboxes and their, C-code based, low-level functions. These toolboxes included Spider v. 1.71, SOM Toolbox v. 2.0, Neural Network Toolbox v. 5.0, Statistics Toolbox v. 5.2 and GEAT toolbox v. 3.5.

Using MTools, the user can build up the (predictive) modeling run, which starts from the selected target data matrix and ends up final model estimates and computed statistical performance indices. The defined model run can involve missing data imputation, data scaling, sine and cosine transformations of discontinuous variables, variable delaying, modeling runs for the selected modeling methods and computation of statistical performance indices for resulting model outputs.

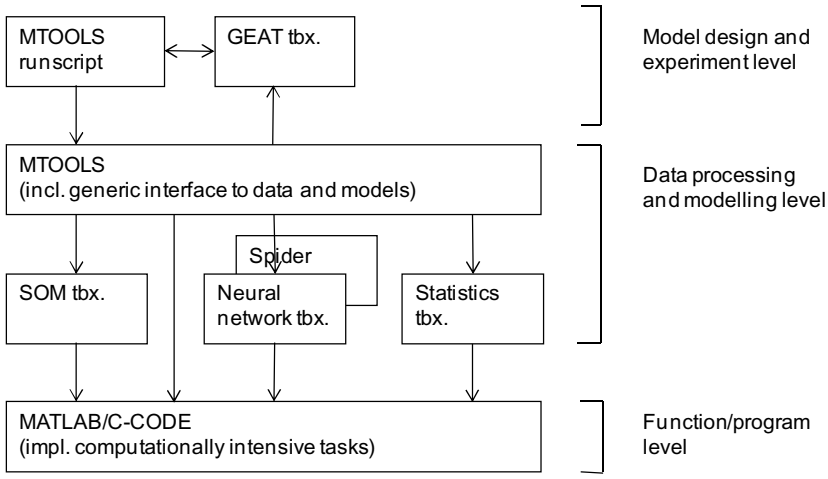


*Figure 8. The architecture of Matlab Mtools modeling library.*

## 4.5 DATA-DRIVEN MODELING APPROACHES

Following the selected computational approach, the data-driven modeling approaches were built and evaluated in the selected application domains. To provide a more compact understanding of the modeling experiments, the stages of data processing and modeling carried out are summarized in each application domain in Table 2.

*Table 2: The main stages of data processing in the predictive data-driven modeling schemes evaluated in this thesis.*

| **Paper** | **I** | **II–III** | **IV** | **V** |
|---|---|---|---|---|
| | **Air quality forecasting** | | **QSARs and chemical grouping** | **ALS-based forest inventory** |
| **Data selection** | Good quality air quality data | Representative years and locations | Based on a set of target chemical compounds | Based on stands of a forest inventory |
| **Handling of missing data** | Missing data replaced using NN (Dixon, 1979) | Missing data replaced using SOM+MLR (Paper I) | Incomplete or constant variables omitted | (no missing data) |
| **Data transformation and scaling** | Variance scaling; sine and cosine transformation of cyclic variables | Variance scaling; sine and cosine transformation of cyclic variables | Variance scaling; logarithmic transformations | Variance scaling; square root, logarithmic transformations and powers |
| **Modeling methods** | MLR, NN, MLR, SOM, MLP | MLP | Sammon's mapping and MLR (PLS, SVM discussed) | SOM, MLP, SVR, k-MSN |
| **Model structure selection** | All the existing input variables selected | GA and multi-objective GA; sensitivity analysis | Based on the existing models or knowledge | Multi-objective GA |
| **Model validation** | Hold-out (artificial gaps) | Hold-out (last year) and cross-validation | Leave-one-out validation | Leave-one-out validation |

Next, the key results and their significance in each application domain are assessed, by reproducing some figures from the original papers. For more throughout application-specific information, the reader is referred to the original research papers.

### 4.5.1 MLP-GA based air quality forecasting

In **Papers II–III**, a data-driven modeling approach based on the standard MLP network was developed and evaluated for the forecasting of urban concentrations of $NO_2$ and $PM_{2.5}$ (Figure 9). The advantage of the proposed MLP-based air quality forecasting (AQF) scheme is, that it does not require exhaustive information on underlying air pollution mechanism and has the ability of modeling non-linear relationships between different predictor variables.

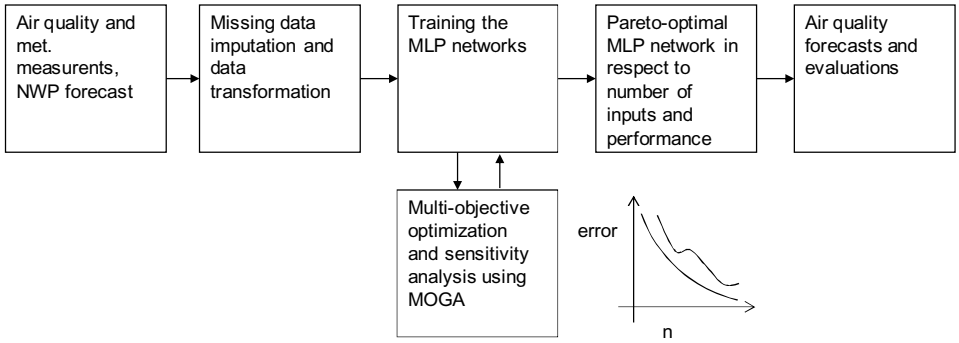


*Figure 9. Computational approach adopted for air quality forecasting.*

The basic idea of the modeling was to establish a site-specific association between emissions, estimated here from time of day inputs, meteorology and air quality. An innovation of the modeling was that the meteorological input data were supplemented with the numerical weather predictions (NWP) provided by the limited area weather forecasting model HIRLAM (Eerola, 2002). Previously, the evaluation of ANN/AQF modeling schemes has been largely performed using the meteorological measurement data instead of the actual NWP data (e.g. Kukkonen et al., 2003). The main stages of the modeling are briefly discussed next.

**Handling of incomplete air quality data**

First the datasets gathered from different sources of information underwent the preprocessing of collected raw datasets in order to achieve harmonized and complete training data. In the data preprocessing, a particular focus was placed on the missing data, which is a common problem with air quality datasets, posing obstacles for the use of standard MLP-modeling.

In **Paper I**, various computational methods were tested using simulated missing data patterns, the methods ranging from mean substitution, interpolation (nearest neighbor, linear and spline) and regression to the standard ANNs (SOM and MLP). It was found that short missing data gaps can be replaced reliably using LIN. The performance of LIN depends, however, strongly on the variable under study (see Figure 1, **Paper I**). Compared to ANNs, the performance of LIN was shown to be better in case of short missing data gaps, the performances being, as expressed using IA, 0.85 for LIN and 0.72–0.81 MLP and SOM (see Table 3, **Paper I**). This is not, however, the case with more complex missing patterns, where the performance of LIN is more degenerated compared to ANNs. This makes a combination of LIN and ANNs attractive for the replacement of missing air quality data sets.

On the basis of the aforementioned observations a hybrid scheme based on a two-stage imputation procedure was tested (**Paper I**, Sect. 2.3.6). The basic idea of the method is to predict the performance of the linear substitution for different gap lengths (hours), from the basis of the average gradient and the exponent $\alpha$, which describes the memory characteristics of time-series. The exponent $\alpha$ (Feder, 1988) is related to Hurst exponent $H$ and fractal dimension $D$, which are commonly used in fractal and spectral analysis for measuring randomness of time-series and roughness of a surface (Voss, 1991). The relationship can be represented using the following MLR model:

$$\exp(d_i) = A_i * \exp(grad_i) + B_i * \exp(\alpha) + C_i \quad (38)$$

where $d_i$ is the index of agreement (marked in the compendium as IA) calculated for different gap lengths $i$, $grad_i$ is the average gradient over the gap length $i$ calculated for every available time point of variable, $\alpha$ is the exponent calculated ignoring the real gaps, and $A_i$, $B_i$ and $C_i$ are regression coefficients for the gaps $i$ calculated from air quality datasets.

The overall accuracy of the proposed MLR model (Eq. 38) was found to be moderately good, the bootstrapped overall performance being 0.81, as expressed using IA, for the air quality variables and for 1–24 hours gap lengths studied (these results are not published). The proposed MLR model can be encapsulated in to the following two-stage imputation scheme:

1. *Stage*

    1. Define a limit value for the imputation performance (e.g. $d_i$ = 0.90) and estimate the critical gap length for a variable using the MLR model (Eq. 38)

       (predict $d_i$ for different gap lengths $i$ based on the calculated variable specific $grad_i$ and the exponent $\alpha$ until $d_i$ decreased below a chosen limit value)

    2. Perform the linear substitution for the missing data gaps under the defined critical gap length

2. *Stage*

    3. Select a multivariate imputation method (e.g. SOM) and perform the missing data imputation for the remaining missing data patterns

The results obtained in the imputation of missing air quality data are in general consistent with the literature (e.g. Latini and Passerini, 2004; Turias et al., 2007; Pisoni et al., 2008), suggesting the good usability of SOM and MLP in recovering the missing data in air quality datasets. The pros and cons of different methods

in respect of different criteria are presented in Table 5, **Paper I**. The benefit of SOM over MLP is, that it is less dependent on the actual location of the missing values, i.e., MLP imputation scheme requires the training of separate networks for different missing data patterns, which can lead to an incoherence between imputed values (Kalteh and Hjorth, 2009). Furthermore, SOM do not generate the values outside the original data range. Therefore, it seems to be safer to rely on SOM.

From the point of ANN/MLP-based AQF models, the recovery of incomplete air quality data is the essential part of the modeling, since the complete/continuous data are required for training and testing networks as demonstrated in **Papers II** and **III**. Alternatively, it is possible to discard all incomplete data rows of the data matrix (e.g. Kolehmainen et al., 2001), but such approach can lead to the loss of a significant amount of information and other potential pitfalls in the modeling.

Finally, it is worthwhile noting, that despite the methods being tested here solely with the air quality data, they could be applied to other type of environmental data with the same structure.

**Selecting MLP model structure**

After the pre-processing of the air quality data, the MLP network was trained to predict hourly $NO_2$ and $PM_{2.5}$ concentrations in time (**Papers II** and **III**). A major concern was the complexity of the MLP network, which was controlled by the selection of appropriate model input variables and the selection of high-level parameters (a number of hidden layers and nodes).

In this thesis, evolutionary selection schemes were employed, which might have some benefits in ANN-based AQF model design. In **Paper II**, the standard GA with one objective was used for evolving appropriate MLP network structure (hidden layer) and relevant input variables simultaneously. It was found that GA is capable of reducing the enhancement of the prediction performance of the model slightly (see Table 2, **Paper II**). A limitation of this approach was, as expected, the high computational demand, which is due to the training of the MLP network for each input subset-structure combination. Furthermore, the ability of single-objective GA to reduce the number of model input variables was found to be limited.

To overcome the problems obtained in **Paper II**, the multi-objective GA (Fonseca and Fleming, 1993; Srinivas and Deb, 1994; Deb, 2004) was tested in **Paper III**. The objective was to minimize the prediction error of the MLP network and the number of model input variables, simultaneously. Furthermore, the sensitivity analysis (SA) of the MLP network proposed by Moody and Utans (1991) was used in conjunction with the multi-objective GA to reduce the computational

burden of the model evolving. Such a SA approach is capable of testing the importance of input variable subsets instead of one input variable at a time. Usually SA is performed by varying only one variable at a time (e.g. Belue and Bayer, 1995), but a clear limitation of such an analysis is that it may not consider an interconnection between other variables and may therefore lead to a misleading analysis (May et al., 2009).

**Accuracy of prediction**

The resulting cross-validated performance statistics showed moderately good general prediction performance for the MLP models studied, IA values ranging between 0.80–0.91 for $NO_2$ and 0.63–0.81 for $PM_{2.5}$ (Table 2, **Paper II**; Tables 2 and 3, **Paper III**). The accuracies obtained are in line with the corresponding accuracies reported in the previous studies (e.g. Kolehmainen et al., 2001; Kukkonen et al., 2003). In the operational set-up (MLP+NWP24, see **Paper III**), NWP/HIRLAM input data were shown to increase the accuracy of the MLP models with the corresponding predictions based solely on the meteorological observations (MLP+MPP00), the IA values increasing from the level of 0.7 to the level of 0.8.

The results obtained for $NO_2$ in **Paper II** were revalidated using the corresponding MLP model trained on the concentration data collected at the Töölö air quality monitoring station, Helsinki during the years 1996–1999 (these results are not published). The model was trained for the years 1996–1999, and tested for the last year 1999.

According to the achieved results, the accuracy obtained using the now-casting MLP model (corresponds to MLP+MPP24, see **Paper III**) was approximately 0.90–0.91, as expressed using IA values. The performance is comparable with the results obtained in **Papers II–III**, which showed the accuracy between 0.86–0.90. The proposed MLP model seems to be capable of predicting the occurrence of highest concentrations to some extent (Figures 10 and 11). This observation is also made by Kukkonen et al. (2003) with the same data. However, this is not the case with the Vallila data, where the models tend to underestimate the highest concentrations throughout (see Figure 4c, **Paper III**).

The true potential of the models is not obvious in the light of calculated statistical performance measures, since the validation schemes used were measuring the performance of the models strictly on an hourly level and did not consider the potential temporal "shift" in the prediction. From the application point of view (air pollution control and early warning), such hourly accuracy is not always necessary, but it is sufficient to derive information on potential episodic situations with lower temporal accuracy (e.g. bihourly, 6-hourly or daily).

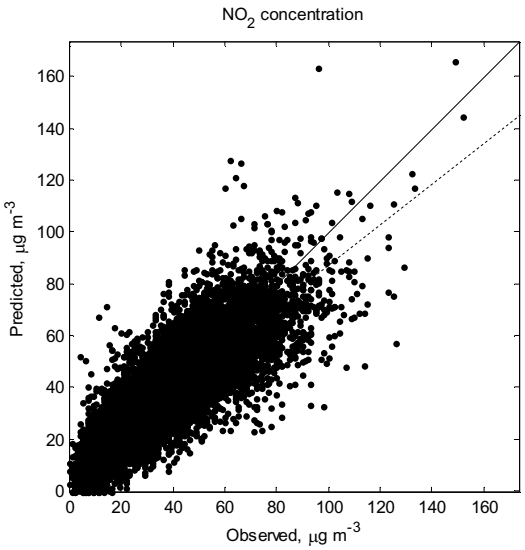


*Figure 10. Measured versus predicted $NO_2$ concentrations at the Töölö station during the year 1999 as obtained with the MLP+MPP24 model. The plot is enhanced with least squares fitting line (dotted) and a line showing perfect fit (solid).*

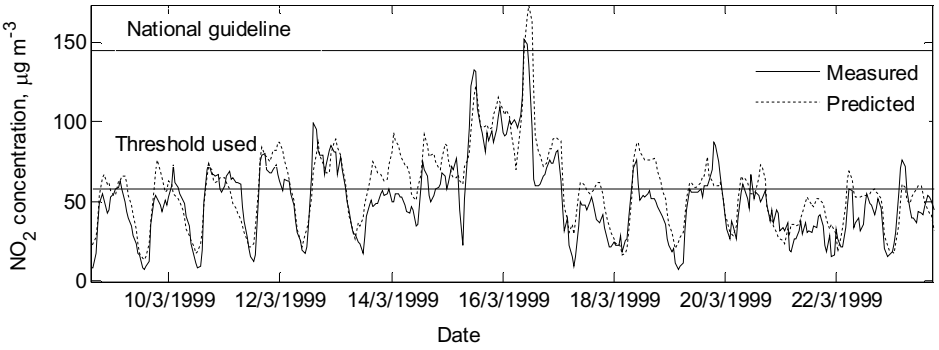


*Figure 11. Hourly time series of the measured and predicted concentrations of $NO_2$ at the station of Töölö from 9 to 23 March 1999 for the MLP+MPP24 model.*

It should be emphasized, that the prediction of infrequent peak pollution events is important to be parallel to the successful modeling the average behavior of a system. In accordance with the obtained results, the limited ability of the MLP models to forecast an occurrence of peak pollution situations is obvious. False alarm rates obtained were in the level of 50% (see Tables 2 and 3, **Paper III**). Underestimation of the highest concentrations was interpreted to be mainly due to the tendency of the MLP model to average over the rare episodic data but,

obviously, also due to the absence of relevant predictor variables in the training data. It is clear, for instance, that some necessary variables are missing in case of long range transportation of $PM_{2.5}$. Further, the meteorological input data gathered from the area-based NWP forecasts and/or area-based monitoring network do not represent the site-specific meteorological conditions sufficiently. Compared with deterministic dispersion modeling (DET), the MLP models, however, seem to exhibit higher site-specific prediction accuracy (e.g. Kukkonen et al., 2003; Rantamäki et al., 2005).

The restriction of the ANN/MLP models still is not being applicable neither for predicting spatial concentration distributions in urban areas nor for evaluating air pollution scenarios for future years (Kukkonen, et al., 2003). This is mainly due to the MLP network being solely based on the experimental data, i.e., it does not imply real physical interactions of the air pollution process, and cannot properly predict/extrapolate pollution situations outside the training data.

Despite the lacks previously discussed, it should be emphasized that the evaluation of the MLP network was performed here based only on restricted training data. Therefore, strong conclusions should be made with care. As a general conclusion, the MLP network can be regarded as a potential tool for the prediction of pollutant concentrations in city hot-spots, however, providing the good quality and representative training data exist. Combined with other modeling tools the MLP-based models can strengthen decisions in operational air quality management and pollution control.

**Recommendations for future work**

Many inherent problems are associated with the MLP network, which include the determination of feasible network structure, the selection of training data, local minima and the curse of dimensionality. These issues were shown in the present study and partly resolved through the GA-based model design. It is, however, obvious that more emphasis should be directed on more advanced methods, such as SVR, wavelets and ensembles, instead of the standard MLP network.

Perhaps more important is, however, the enhancement of the capabilities of ANN network to learn extreme and spatially dependent characteristics of urban air quality. The lack of the ANN model trained strictly on using site-specific data is, that it may easily overfit the data when trying to reproduce extreme values (Foresti et al., 2010). Enhancements could be found for instance from Extreme Value Theory (EVT), which is a branch of statistics for analyzing the tail behavior of a distribution (Gumbel, 1958; Beirlant et al., 2004). Many interesting statistical modeling approaches, which adopt the principles of EVT can be found (e.g. Easteo, 2007). Another potential research direction is to combine ANNs

with geostatistical modeling methods (such as interpolation methods). Geostatistical methods can be used to produce more representative spatial training data for ANN modeling, thus enabling also predictions for spatial concentration distributions required in air quality management (e.g. Attore et al., 2007; Foresti et al., 2010; Foresti et al., 2011).

### 4.5.2 Novel QSAR and chemical grouping approach

In **Paper IV**, QSAR modeling and the chemical grouping approach based on a combination of Sammon's mapping and MLR models was evaluated for the characterization of unknown physicochemical properties and (eco)toxicity of a set of target chemical substances within the REACH regulation (Figure 12). The basic goal was to explore the similarity between the target chemical substances and the set of more information rich reference chemical substances (see Table 3, **Paper IV**) in order to apply the read-across within the identified chemical substance categories. The proposed approach can be considered as a sort of semi-quantitative prediction method for the interpolation of existing information from a set of reference chemical substance to a target chemical substance.

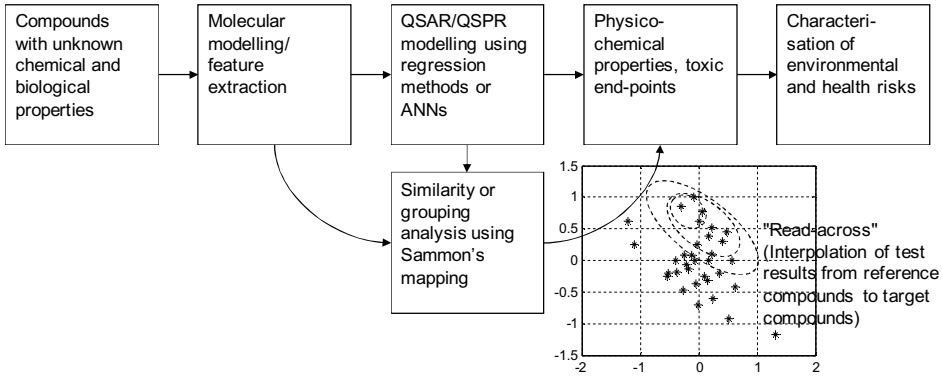


*Figure 12. (Q)SAR approach based on QSAR predictions and Sammon's mapping (modified from Figure 2, **Paper IV**).*

A particular issue with the QSAR modeling and chemical grouping was that of reducing the dimensionality of multivariate descriptor data. This was required for better discovery and visualization of chemical groups for target substance chemicals, which is essential for the adoption of the read-across for filling-in data gaps in REACH. Previously, the dimensionality reduction has been usually performed using PCA or related approaches (e.g. Wold et al., 1987; Geladi and Kowalski, 1986). However, PCA in its basic form cannot provide a compact and

easily understandable picture about the chemical clusters in 2D space. To overcome the previous lack, Sammon's mapping was used to explore the similarity of the chemical substances in two feature dimensions. The basic goal was to determine whether changes in chemical structure lead to any marked shift in chemical and (eco)toxicological properties.

The modeling approach adopted followed the principles of the (Q)SAR analysis, which starts with the computation of molecular descriptors of a set of molecules with limited structural variability and known responses. These data are then fed into the basis of the data-driven analysis. In this thesis, DRAGON-based structural descriptors were computed. These DRAGON data were supplemented with physicochemical and (eco)toxic data produced by the simple MLR models (e.g. Papa et al., 2005; Gramatica et al., 2007). Sammon's mapping was then used to discover chemical substance groups in that high dimensional structural and property/activity descriptor space (Figure 13).

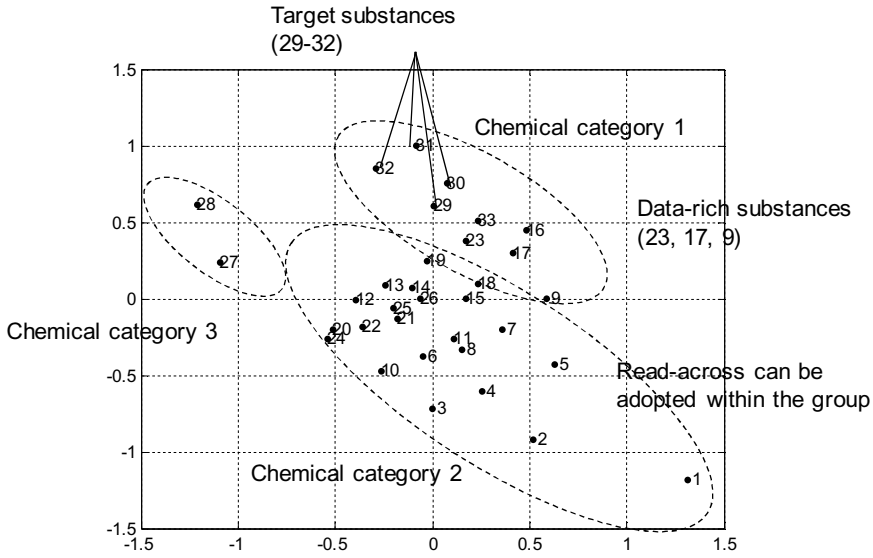


*Figure 13. Principle of the Sammon's mapping-based chemical grouping (reproduced with recalculation from Figure 8,* ***Paper IV****).*

**Regulatory acceptance of the modeling**

The obvious benefit of the approach is the ability of Sammon's mapping to visualize the patterns and analogies for a set of target chemicals in two dimensions. This could be beneficial from the application point of view (the REACH regulation), as it facilitates the visual interpretation of the results.

Whenever the (Q)SAR models are used for regulatory purposes, the models should however fulfill regulatory validation principles, i.e., the sufficiency of information for making a particular regulatory decision (Tichý and Rucki, 2009). According to OECD (2007) the following validation principles are considered useful for the assessment of regulatory models:

- A defined endpoint
- An unambiguous algorithm
- A defined domain of applicability
- Appropriate measure of goodness-of-fit
- Robustness and predictive ability
- Mechanistic interpretation, if possible

The previous criteria are complementary to the REACH regulation, in which (i) the relevance, (ii) the reliability and (iii) the adequacy of the models is required (Worth et al., 2007; ECHA, 2010). The relevance, the reliability and the adequacy refer to the appropriateness of the prediction in relation to the information needed for the regulatory purpose, the inherent quality i.e. validity of the model and its applicability and the sufficiency of information for making a particular regulatory decision.

The mechanical interpretation of the proposed analysis can however be impossible or impractical. Further, the approach tends to reflect more the overall structural similarity (cf. DRAGON descriptors) than the biological activity when it is based on a high number of complex structural molecular descriptors. To overcome these deficiencies, the approach was compensated by mechanical-based selection of key descriptor variables, which were related to the basic structural information and well-defined physico-chemical and biological properties of the substances under study.

**Challenges with the modeling**

Taken overall, the approach produces the semi-quantitative results, which still contain highly supportive information for the regulatory purpose. It should be emphasized, that the (Q)SAR predictions are valid only within the domain of training/calibration data, and whenever the models are used for external predictions, some basis in physical reality is required (e.g. Johnson, 2008).

Limitations remain, especially when it comes to the assessment of more complex health related endpoints (e.g. Schultz et al., 2003; Cronin and Worth, 2008). Health endpoints are many times reflected by differences in mode of action between chemical classes and differences in toxicokinetics and toxicodynamics between species. On the other hand, it has been found that the most critical aspects with respect to the development of valid QSARs seems not to be model

concepts, but high uncertainties and inconsistency in the data used for calibrating QSARs (e.g. Thomsen, 2001).

As a recommendation for further work, more emphasis should be placed on building up extensive QSAR databases, which enable further testing and evaluation of new data-driven modeling concepts, especially, for the modeling of a complex biological activity. Further toxicological data are required, which could be released through initiatives related to the REACH regulation (Cronin and Worth, 2008).

### 4.5.3 ANN-GA based forest inventory modeling

In **Paper V**, three ANN models, namely MLP, SVR and SOM, were evaluated for the prediction of stem volumes ($m^3ha^{-1}$) of Norway spruce, Scots pine, and deciduous trees at the field plot and forest stand levels using the features calculated from the ALS data and aerial photographs. Previously, the ALS-based prediction of continues forest attributes has been mainly based on conventional parametric and non-parametric regression methods (e.g. Naesset et al., 2005; Packalén and Maltamo, 2006). Although the relatively good prediction accuracies have been achieved, the methods have faced with many potential shortcomings originated from the characteristics of the ALS data. Therefore, new, more powerful methods capable of handling high dimensional RS data, which may contain irrelevant, correlated and noisy variables, are required for ensuring reliable forest attribute estimates.

The ANN methods were compared with the corresponding *k*-MSN method (Packalén and Maltamo, 2006). According to the achieved LOO performance statistics (Table II and III, **Paper V**) ANN modeling can be regarded as an appropriate approach for the ALS-based forest inventory resulting in high prediction performance (see Figure 14). The accuracies obtained with different ANN models were, in terms of IA, between 0.71–0.96 in plot level and 0.84–0.98 in stand level, which can be considered to be sufficient for forest inventory purposes. The achieved performances are in general comparable with the corresponding accuracies obtained in related studies (e.g. Packalén and Maltamo, 2007).

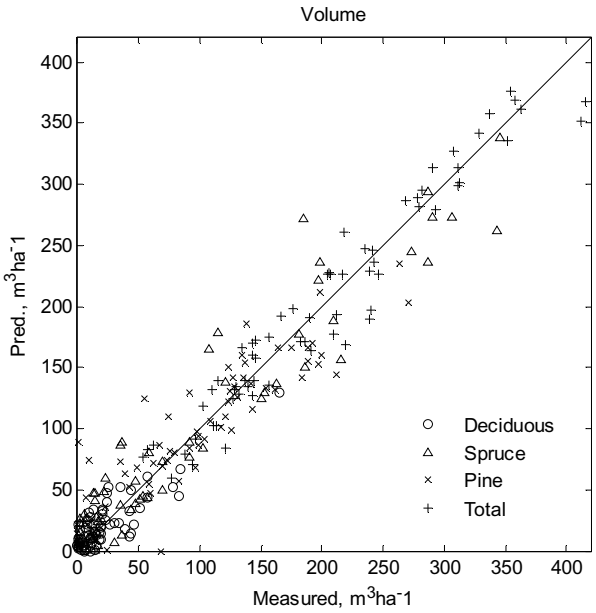


*Figure 14. Measured versus predicted stand volumes by tree species and total volume as obtained with MLP (reproduced from Figure 3, **Paper V**).*

The highest prediction accuracies were obtained with the MLP and SVR methods. The accuracy of SOM was found to be degenerated compared to the corresponding accuracies achieved using MLP and SVR. A problem observed with the direct application of MLP and SVR was that they produced the negative values outside the original data range. However, it might be possible to avoid this problem with aid of specific data transformation methods, which were out of scope in this thesis.

To sum up, a relatively limited dataset was used to evaluate and compare the ANN methods in this thesis. Consequently, further evaluation of the ANN methods is required using more extensive ALS data. This could show better the potential of an ANN-based forest inventory analysis.

In addition to the novel application of ANN modeling methods, the major innovation of the study was to show the usability of the multi-objective GA for selecting appropriate ALS variables in ALS-based forest inventory. Previously, the selection of the ALS features has often been based on time-consuming trial-and-error experiments using iterative insertions and deletions, more advanced but local heuristic procedures or filter based distance metrics such as Minkowski distances (e.g. Packalén and Maltamo, 2006; Packalén and Maltamo, 2007; Peuhkurinen et al., 2008). The use of global search procedures has been largely omitted, expect for some recent studies (e.g. Holopainen et al., 2010). In general,

the multi-objective GA-based variable selection method was found to be well-suited for the selection of appropriate ALS variables for species-specific forest inventory models. Further, the combination of *k*-MSN and multi-objective GA was shown to be powerful alternative for the input variable selection. A particular benefit of such the approach is its ability for global of exploration of the ALS feature subsets.

# 5 Summary and conclusions

In this thesis, the usability of the modern computational methods and related data-driven modeling schemes was evaluated in solving the predictive modeling problems associated with environmental management decision-making. The selected case studies included (i) the forecasting of urban airborne pollutant concentrations, (ii) the prediction of physicochemical and biological properties of a set of target chemical substances using quantitative structure-activity relationships (QSARs) and chemical grouping, as well as (iii) the prediction of species-specific forest attributes using airborne laser scanning (ALS) data and other RS data.

The investigation was mainly based on experimental model design and evaluation work, with an examination of the external validity using a comparison of model output with the experimental data. The main modeling methods investigated included artificial neural networks (ANNs) and related methods, among them multi-layer perceptron (MLP), support vector regression (SVR) and self-organizing map (SOM) and Sammon's mapping. In addition to the novel applications of the modeling methods, the main innovation of the thesis was to demonstrate the usability of genetic algorithm (GA) based optimization schemes for determining appropriate model input variables and structure in the application domains studied. Even though approaches based on the use of GA have been presented in the related fields of environmental modeling, they have not been previously applied to this extent in the selected applications.

First, the experiments were conducted with air quality data. In the first stage, the computational approaches based on MLP and SOM were developed and tested for imputing missing data in air quality datasets. The results obtained show that MLP and SOM are well-suited and relatively accurate methods for missing data imputation. The advantage of SOM over MLP is that it is less dependent on the actual location of the missing data and thus it seems to be safer rely on it. Moreover, it is shown that the performance of linear substitution in respect to the length of gaps can be estimated separately for each variable of air quality by calculating a gradient and the exponent $\alpha$, the latter describing the memory characteristics of the variable time-series. This relationship can be encapsulated into the hybrid method proposed in the thesis.

In the second stage of the air quality studies, the MLP-based modeling schemes were built and evaluated for the forecasting of airborne pollutant concentrations in city hot-spots. This part of the work aimed to strengthen the previous

knowledge about the usability and accuracy of MLP-based models in air quality forecasting and in particular, to yield new information about the accuracy of MLP models in the operational condition where numerical weather prediction (NWP) data are available. This is important since the evaluation of MLP-based models has been largely made using the meteorological observations instead of the actual NWP data. In addition to the novel applications of MLP, the main objective was to evaluate the usability of novel GA-based optimization schemes for selecting the structure and input variables of the MLP model needed in air quality forecasting.

The results of the MLP-based air quality forecasting in general show moderately good prediction performance for airborne concentrations of $NO_2$ and $PM_{2.5}$. In addition, it is shown that the operational accuracy of the MLP network can be enhanced using the forecasts of the NWP model as the model input. The performance of MLP is, however, degenerated in the course of peak pollution episodes where pollutant concentrations reach their highest values. This is a clear shortcoming from the urban air pollution control point of view and requires thus more attention in the future. Throughout the air quality studies, GA-based optimization schemes are shown to be well-suited for designing the MLP-based air quality models. The major observation from the experiments with GA is that the multi-objective GA combined with the sensitivity analysis (SA) of the MLP network provides a computationally powerful approach for the input variable selection, which could be useful also in other related problem domains.

In the second stage of the thesis, novel chemical grouping approach, based on the combination of Sammon's mapping and simple regression-based quantitative structure-activity relationships (QSAR) models was developed and applied for characterizing unknown physicochemical and biological properties of chemical substances under the information requirements of the REACH (2006/1907/EC). The results and observations show that Sammon's mapping is a powerful method for discovering and visualizing chemical substance groups and their internal physicochemical and biological analogies from the chemical descriptor data. The advantage of Sammon's mapping is its capability of compressing highly multidimensional and collinear descriptor data into visually interpretable two dimensions. Such a chemical grouping approach is expected to be a suitable read-across approach for predicting physicochemical properties, human health effects and environmental effects from the data of information rich reference substances in REACH. This could then reduce the need of expensive and laborious laboratory testing, where every substance is tested for every endpoint.

Lastly, ANN modeling methods were presented in airborne laser scanning (ALS) based forest inventory, where they have not been previously applied to this

extent. Three ANN-based ALS models, namely SOM, MLP and SVR, were compared to the *k*-MSN method previously adopted in ALS-based forest inventory. In addition, the multi-objective GA was used to select appropriate model input variables among the large number of potential ALS-based descriptors. The results obtained show that MLP and SVR produce reliable estimates for species-specific forest attributes. The accuracy of SVR and MLP is comparable with the corresponding accuracy obtained with the *k*-MSN method. In addition, it is shown that the performance of the models can be enhanced using the GA-based optimization scheme.

In accordance with the results obtained with ANNs, MLP and SVR produced highest prediction performances. The problem of MLP network is, however, the risk of over-fitting, which makes the use of SVR more favorable option for the modeling. In addition, multi-objective GA was shown to be a respectable method for selecting appropriate input variables for the modeling. In that respect, it seems reasonable to use GA with SA or simpler predictor (such as linear regression) as a filter approach and then train a more complex non-linear ANN model on the resulting variables. This also concurs with the results reported in the literature (e.g. Guyon and Ellissef, 2003).

Overall, the results and observations of this thesis suggest that the modern computational approaches studied are well-suited for solving complex prediction problems of environmental management, however, providing that good quality and representative datasets exist in problem domains. More precisely, it is shown that the computational approaches studied have great capability of:

- Enhancing the quality of environmental data for further data processing and modeling in environmental management.
- Producing sufficient and in-time predictions on complex environmental systems and processes required by decision-makers.
- Optimizing the structure of data-driven models in order to achieve computationally more powerful, predictive and interpretable models required by applications of environmental management.
- Replacing existing laborious and expensive field and laboratory testing procedures currently used in various assessment procedures of environmental management.
- Discovering interpretable clusters/patterns from large environmental datasets, which can be used to suggest hypothesis and to support conclusions in environmental management.

The results of this thesis can be used in the development of new, more accurate data-driven modeling schemes required by applications of environmental management, as well as the scientific base for further studies. In the future, further development of predictive modeling is required, especially, in respect to

the modeling of rare and spatially dependent processes, as shown in this thesis through the prediction of infrequent peak pollution episodes. The combination of the modern data-driven modeling methods and geostatistical modeling methods is thus one potential research direction. In addition, more emphasis should be placed on improving the mechanistic interpretation of the data-driven models in order to improve their (regulatory) acceptance. This requires the development of hybrid modeling approaches where physical information about underlaying phenomenon is encapsulated at some level into the data-driven models.

# *References*


Abebe, A.J., Solomatine, D. P., and Venneker, R.G.W. (2000) Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrological Sciences Journal 45(3), 425-436.

Asikainen, A. (2006) Use of Computational Tools for Rapid Sorting and Prioritising of Organic Compounds Causing Environmental Risk and Cytochrome P450 Activity. Kuopio University Publications C. Natural and Environmental Sciences 191.

Atkinson, P.M. and Tatnall, A.R.L. (1997) Neural networks in remote sensing, Introduction. International Journal of Remote Sensing, 18(4), 699–709.

Attore F., Alfo M., Sanctis M., Francesconi F. and Bruno F. (2007) Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale. International Journal of Climatology 27, 1825–1843.

Avouris, N.M. and Page, B. (1995) Environmental Informatics: Methodology and Applications of Environmental Information Processing, Kluwer Academic Publishers, Dodrecht.

Baklanov, A., Rasmussen, A., Fay, B., Berge, E. and Finardi, S. (2002) Potential and shortcomings of numerical weather prediction models in providing meteorological data for urban air pollution forecasting. International Journal of Water, Air and Soil Pollution 2, 43–60.

Barry, S. and Elith, J. (2006) Error and uncertainty in habitat models. Journal of Applied Ecology 43, 413–423.

Belue, L.M. and Bauer, K.W. (1995) Determining Input Features for Multilayer Perceptrons. Neurocomputing 7, 111–121

Berthouex, P.M. and Brown, L.C. (2002) Statistics for Environmental Engineers, 2nd Edition. Lewis Publishers, Boca Raton, FL, USA.

Bezdek, J.C. (1994) What is computational intelligence? In Zurada, J., Marks, J.M., and Robonson, C. (Eds.), Computational Intelligence: imitating life, Piscataway, IEEE Press, 1–12.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) Statistics of Extremes: Theory and Applications, Wiley Press.

Bishop, C. (1995) Neural Networks for Pattern Recognition, Clarendon Press, Oxford.

Box, G.E.P. and Jenkins, G.M. (1970) Time Series Analysis, Forecasting and Control. San Francisco, CA.

Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2), 121–167.

Bäck, T. (1996) Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming and Genetic Algorithms. Oxford University Press.

Canu, S. and Rakotomamonjy, A. (2001) Ozone peak and pollution forecasting using support vectors. In IFAC Workshop on environmental modelling, 22–23 August, 2001, Yokohama.

Castillo, P.A., Arenas, M.G., Castillo-Valdivieso, J.J., Merelo, J.J., Prieto, A. and G. Romero, G. (2002) Artificial neural networks design using evolutionary algorithms. Proceedings of the Seventh World Conference on Soft Computing, 2002.

Chelani, A. (2010) Prediction of daily maximum ground ozone concentration using support vector machine. Environmental Monitoring and Assessment 162, 169–176.

Chen, K.-Y. and Wang, C.-H. (2007) Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management, 28, 215–226.

Cherrkassky, V. and Mulier, F. (1998) Learning from Data: Concepts, Theory and Methods. New York: Wiley.

Cherkassky, V. and Ma, Y. (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1), 113–126.

Cherkassky, V., Karnopolsky, V., Solomantine, D.P. and Waldes, J. (2006) Computational intelligence in earth sciences and environmental applications: Issues and challenges. Neural Networks, 19(2), 113–121.

Commission of the European Communities (2006) Regulation (EC) No. 1907/2006 of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Available at http://europa.eu.int/comm/enterprise/reach/ index_en.htm (Accessed 7 August 2007).

Cronin, M.T.D. and Worth, A.P. (2008) (Q)SARs for predicting effects relating to reproductive toxicity. QSAR and Combinatorial Science 27, 91–100.

Daescu, D.N. (2009) Sensitivity analysis methods in air quality models. In Hanharan, G. (Ed.) Modelling of Pollutants in Complex Environmental Systems. ILM Publications: Hertfordshire.

Deb, K. (2004). Multi-Objective Optimization using Evolutionary Algorithms. Chichester: Wiley.

Dixon, J.K. (1979) Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics 10 (SMC-9), 617–621.

Dowdy, S., Wearden, S., and Chilko, D. (2004). Statistics for Research. 3rd Edition. John Wiley & Sons, Hoboken, NJ, USA.

Drucker, H., Burges, C.J.C, Kaufman, L., Smola, A. and Vapnik. V. (1997) Support vector regression machines. In M. C. Mozer, M. I. Jordan, T. Petsche, editors, Advances in Neural Information Processing Systems, volume 9, p. 155. The MIT Press.

Easteo, E. (2007) Statistical models for dependent and non-stationary extreme events. PhD thesis, Lancester University.

Eerola, K. (2002) The operational HIRLAM at the Finnish Meteorological Institute. HIRLAM Newsletter 41, 19–24.

Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York.

Emmanouilidis, C., Hunter, A., Macintype, J. and Cox, C. (2001) A multi-objective genetic algorithm approach to feature selection in neural and fuzzy modeling. Evolutionary Optimization 3, 1–26.

Engelbrecht, A.P. (2007) Computational Intelligence: An Introduction, 2nd edition. John Wiley & Sons, West Sussex, England.

Esbensen, K.H. (2002) Multivariate Data Analysis – In Practice: An Introduction to Multivariate Data Analysis and Experimental Design. 5th Edition, CAMO Process AS, Oslo, Norway, p. 598.

European Centre for Ecotoxicology and Toxicology of Chemicals (1998) QSARs in the Assessment of Environmental fate and Effects of Chemicals, ECETOC Technical Report, No. 74, Brussels, Belgium.

European Chemicals Agency (2010) REACH guidance, Chapter R.6: QSARs and grouping of chemicals. Available at http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r6_en.pdf?vers=20_08_08 (Accessed 7 January 2010)

Fayyad, U. (1996) Data mining and knowledge discovery: Making sense of data. IEEE Expert Intelligent Systems and Their Applications 11, 20–25.

Feder, J. (1988) Fractals. Plenum Press, New York.

Fletcher, D., MacKenzie, D. and Villouta, E. (2005) Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. Environmental and Ecological Statistics 12, 45–54.

Fogel, D.B. (1995) Review of "Computational intelligence: imitating life". IEEE Transactions on Neural Networks 6, 1562–1565.

Fogel, D.B. (2006) Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. 3rd Edition. IEEE Press, Piscataway, NJ.

Foresti, L., Pozdnoukhov, Tuia, D. and Kanevski, M. (2010) Extreme precipitation modeling using geostatistics and machine learning algorithms. In Atkinson, P.M. and Lloyd, C. D. (Eds.), geoENV VII, Geostatistics for Environmental Applications, Springer, New York.

Foresti, L., Tuia, D. and Kanevski, M. (2011) Learning wind fields with multiple kernels. Stochastic Environmental Research and Risk Assessment 25, 56–66.

Fonseca, C.M. and Fleming, P.J. (1993) Genetic algorithms for multi-objective optimisation: formulation, discussion and generalisation. In Forrest, S. (Ed.), Proceedings of the 5th International Conference on Genetic Algorithms, San Mateo, California, pp. 416–423. Morgan Kaufmann.

Fonseca, C.M. and Fleming, P.J. (1995) An overview of evolutionary algorithms in multiobjective optimization. Evolutionary Computation 3, 1–16.

Fox, D.G. (1981) Judging air quality model performance: A summary of the AMS Workshop on Dispersion Model Perfomance. Bulletin of the American Meteorological Society 62, 599–609

Gardner, M.W. and Dorling, S.R. (1998) Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

Geladi, P. and Kowalski, B. (1986) Partial least-squares regression: A tutorial. Analytica Chimica Acta 185, 1–17.

Gentili, S., Magnaterra, L. and Passerini, G. (2003) An introduction to the statistical filling of environmental data time series. In Latini and Passerini (Eds.), Handling Missing Data: Applications to Environmental Analysis, Billerica, MA: WIT Press.

Geman, S., Bienenstock, E. and Doursat, R. (1992) Neural networks and the bias/variance dilemma. Neural Computation 4, 1–58.

Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, MA.

Gramatica, P., Giani, E. and Papa, E. (2007) Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, Journal of Molecular Graphics and Modelling 25 , 755–766.

Green, D.G. and Klomp, N.I. (1998) Environmental informatics – a new paradigm for coping with complexity in nature. In Stadish, R. et al. (Eds.). Complexity Between the Ecos: From Ecology to Economics. Complex Systems '98, pp. 32–39.

Gumbel, E.J. (1958) Statistics of Extremes, Columbia University Press.

Guyon I. and Elisseeff A (2003) An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, 1157–1182.

Hagan, M.T. and Menhaj, M. (1994) Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 5, 989–993.

Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition. Morgan Kauffmann Publishers, San Francisco, CA.

Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, E. and Streich, M. (1963) The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. Journal of the American Chemical Society 85, 2817–2824.

Hand, D., Mannila, H. and Smyth, P. (2001) Principles of Data Mining, Cambridge (MA), MIT Press.

Hanrahan, G. (2009) Modelling of Pollutants in Complex Environmental Systems. ILM Publications, Hertfordshire.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York

Haupt, S. E., Pasini, A. and Marzban, C. (2008) Artificial Intelligence Methods in the Environmental Sciences. Springer, Berlin.

Haykin, S. and Principe, J.C. (1998) Making sense of complex world. IEEE Signal Processing Magazine, 66-81.

Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, 2nd Edition. Prentice-Hall, Upper Saddle River, NJ.

Holland, J.H. (1975) Adaptation in natural and artificial systems. University of Michigan press, Ann Arbor.

Holopainen, M., Haapanen, R., Karjalainen, M., Vastaranta, M., Hyyppä, J., Yu, X., Tuominen, S., and Hyyppä, H. (2010) Comparing accuracy of airborne laser scanning and TerraSAR-X radar images in the estimation of plot-level forest variables. Remote Sensing 2, 432–445.

Horn, J., Nafploitis, N. and Goldberg, D.E. (1994) A niched Pareto genetic algorithm for multiobjective optimization," in Proceedings of the First IEEE Conference on

Evolutionary Computation, Z. Michalewicz, Ed. Piscataway, NJ: IEEE Press, 1994, pp. 82–87.

Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Jakeman, A.J., Letcher, R.A., Norton, J.P. (2006) Ten iterative steps in development and evaluation of environmental models. Environmental Modeling and Software 21, 602–614.

Johnsson, S.R. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). Journal of Chemical Information and Modeling 48, 25–26.

Jolliffe, I.T. (2002) Principal component analysis. 2nd edition. Springer-Verlag, New York.

Juhos, I., Makra, L. and Tóth, B. (2008) Forecasting of traffic origin NO and $NO_2$ concentrations by support vector machines and neural networks using principal component analysis. Simulation Modelling Practice and Theory 16, 1488–1502.

Kalteh, A.M. and Hjorth, P. (2009) Imputation of missing values in a precipitation – runoff process database. Hydrology Research 40, 420–432.

Karppinen, A., Joffre, S.M., Kukkonen, J., (2000) The refinement of a meteorological preprocessor for the urban environment. International Journal of Environment and Pollution 14, 565–572.

Kohavi, R. and John, G.H. (1997) Wrappers for feature subset selection. Artificial Intelligence, 97, 273–324.

Kohonen, T. (2001) Self-Organizing Maps, 3nd, extended edition. Springer, Berlin.

Kolehmainen, M., Martikainen, H., Hiltunen, T. and Ruuskanen, J. (2001) Neural networks and periodic components used in air quality forecasting. Atmospheric Environment, 35, 815–825.

Kolehmainen, M. (2004) Data exploration with self-organizing maps in environmental informatics and bioinformatics. Kuopio University Publications C. Natural and Environmental Sciences 167.

Krasnopolsky, V.M. and Chevallier, F. (2003) Some neural network applications in environmental sciences. Part II: Advancing computational efficiency of environmental numerical models. Neural Networks 16, 335–348.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R. and Cawley, G. (2003) Extensive evaluation of neural network models for the prediction of $NO_2$ and $PM_{10}$ concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 37, 4539–4550.

Kukkonen, J., Pohjola, M., Sokhi, R.S., Luhana, L., Kitwiroon, N., Rantamäki, M., Berge, E., Odegaard, V., Slordal, L.H., Denby, B., Finardi, S. (2005) Analysis and evaluation of selected local-scale $PM_{10}$ air pollution episodes in four European cities: Helsinki, London, Milan and Oslo. Atmospheric Environment 39, 2759–2773.

Laasasenaho, J. (1982) Taper curve and volume function for pine, spruce and birch, Communicationes Instituti Forestalis Fenniae 108, 1–74.

Latini, G. and Passerini, G. (2004) Handling missing data; applications to environmental analysis. WIT Press, Southampton, Boston.

Leeuwen, C.J. and Vermeire, T.G. (2007) Risk Assessment of Chemicals: An Introduction. 2nd Edition. Springer: Dordrecht, The Netherlands.

Liu, H. and Motoda, H. (1998) Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Massachusetts.

Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. Wiley, New York.

Ljung, L. (1999) System Identification: Theory for the User. 2nd Edition. Prentice-Hall, Englewood Cliffs, NJ.

Lu, W.-Z., Wang, W., Leung, A., Lo, S.-M., Yuen, R., Xu, Z. and Fan, H. (2002) Air pollutant parameter forecasting using support vector machines. In Proceedings of the 2002 International Joint Conference on Neural Networks, 630–635, May 12-17, 2002, Havaij.

Lu, W.-Z. and Wang, W.-J. (2005) Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. Chemosphere 59, 693–701.

Magnaterra, L., Passerini, G. and Tascini, S. (2003). Data validation and data gaps in environmental time series. In Latini and Passerini (Eds.), Handling Missing Data: Applications to Environmental Analysis, Billerica, MA: WIT Press.

Maier, H.R., Ascough, J.C., Wattenbach, M., Renschler, C.S., Labiosa, W.B. and Ravalico, J.K. (2008) Uncertainty in environmental decision making: issues, challenges and future directions. In Jakeman, A.J. et al (Eds.), Environmental Modeling, Software and Decision Support – State of the Art and New Perspectives. Elsevier: Amsterdam.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A. and Kangas, J. (2006) Nonparametric estimation of plot volume using laser scanning, aerial photography and stand register data. Canadian Journal of Forest Research 36, 426–436.

Man, K.F., Tang, K.S. and Kwong, S. (1999) Genetic Algorithms: Concepts and Designs. Springer-Verlag: London.

May, R.J., Maier, H.R., Dandy, G.C. (2009) Developing artificial neural networks for water quality modeling and analysis. In Hanharan, G. (Ed.) Modelling of Pollutants in Complex Environmental Systems. ILM Publications: Hertfordshire.

McCune, B. (1997) Influence of noisy environmental data on canonical correspondence analysis. Ecology 78(8), 2617–2623.

Michaelsen, J. (1987) Cross-validation in statistical climate forecast models. Journal of Climate and Applied Meteorology 26, 1589–1600.

Mileva-Boshkoska, B. and Stankovski, M. (2007) Prediction of missing data for ozone concentrations using support vector machines and radial basis networks. Informatica 31, 425-430.

Miller, G.F, Todd, P.M. and Hegde, S.U. (1989) Designing neural networks using genetic algorithms. In Schaer, J.D. (Ed.), Proceedings of the Third International Conference on Genetic Algorithms, pages 379-384, San Mateo, 1989.

Mitchell, T.M. (1997) Machine Learning. McGraw-Hill, New York.

Monteiro, A, Lopes, M., Miranda, A.I., Borrego, C., Vautard, R. (2005).Air pollution forecast in Portugal: a demand from the new air quality framework directive. International Journal of Environment and Pollution 5, 1–9.

Moody, J. and Utans, J. (1991) Principled architecture selection for neural networks: application to corporate bond rating predictions. In Moody, J., Hanson, S.J. and Lippmann, R.P. (Eds.), Proceedings of Advances in Neural Information Processing Systems. Morgan Kaufmann, San Mateo, CA, pp. 683–690.

Mouer, M. (1987) Nearest neighbour inference for correlated multivariate attributes. In Proc. IUFRO forest growth modelling and prediction conference, USDA forest service, general technical report NC-120, Minneapolis, 23–27 August.

Mouer, M. and Stage, A. R. (1995) Most similar neighbour: An improved sampling inference procedure for natural resource planning. Forest Science 41, 337-359.

Mujunen, S.-P. and Minkkinen, P. (1996) PCA and PLS methods applied to exotoxical data: ecobalance project. Journal of Chemometrics 10, 411–424.

Næsset, E., Gobakken, T., Holmgren, J., Hyyppä, H., Hyyppä, J., Maltamo, M., Nilsson, M., Olsson, H., Persson, Å. and Söderman, U. (2004) Laser scanning of forest resources: The Nordic experience, Scandinavian Journal of Forest Resources 19, 482–499,

Næsset, E., Bollandsås, O.M. and Gobakken, T. (2005) Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. Remote Sensing of Environment 94, 541–553.

Norazian, M.N., Shukri, Y.A., Azam, R.N., Mustafa Al Bakri, A.M. (2008) Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 34, 341–345.

Nunnari, G., Nucifora, A.F.M. and Randieri, C. (1998) The application of neural techniques to the modelling of time-series of atmospheric pollution data. Ecological Modeling 111, 187–205.

Nunnari (2003) Modelling air pollution time-series by using wavelet functions and genetic algorithms. Soft Computing 8, 173–178.

OECD (2007) OECD Principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models, Available at http://www.oecd.org/document/23/0,2340,en_2649_34365_33957015_1_1_1_1,00.html (Accessed 6 August 2007)

Osowski, S. and Garanty, K. (2007) Forecasting of the daily meteorological pollution using wavelets and support vector machine. Engineering Applications of Artificial Intelligence 20, 745-755.

Packalén, P. and Maltamo, M. (2006) Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. Forest Science 52, 611–622,

Packalén, P. and Maltamo, M. (2007) The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment 109, 328–341.

Packalén, P. (2010) Using airborne laser scanning data and digital aerial photographs to estimate growing stock by tree species. Dissertationes Forestales 77, 41 p.

Page, B. and Hilty, L.M. (1995) Umweltinformatik – Informatikmethoden für Umweltschutz und Umweltforchung, Handbuch der Informatik, 2nd Edition, München.

Page, B. and Raustenstrauch, C. (2001) Introduction to environmental informatics systems, In Raustenstrauch, C. and Patig, S. (Eds.), Environmental Information Systems in Industry and Public Administration, Idea Group Publishing, London.

Pal, N.R. and Pal, S. (2002) Computational intelligence for pattern recognition. Internal Journal of Pattern Recognition and Artificial Intelligence 16, 773–779.

Papa, E., Fulvio, V. and Gramatica, P. (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in pimephales promelas (fathead minnow). Journal of Chemical Information and Modelling 45, 1256–1266.

Peuhkurinen, J., Maltamo, M. and Malinen, J. (2008) Estimating species-specific diameter distributions and saw log recoveries of boreal forests from airborne laser scanning data and aerial photographs: A distribution-based approach. Silva Fennica 42(4), 625–641.

Piegorsch, W.W and Bailer, A.J. (2005) Analyzing Enviromental Data. John Wiley and Sons: Chichester, England.

Pisoni, E., Pastor, F. and Volta, M. (2008) Artificial neural networks to reconstruct incomplete satellite data: application to the Mediterranean sea surface temperature. Nonlinear Processes in Geophysics 15, 61-70.

Rantamäki, M, Niska, H., Kauhaniemi, M., Kolehmainen, M., Kukkonen, J. and Karppinen, A., (2005) An evaluation of a deterministic modelling system and a neural network model for forecasting the concentrations of PM2.5. AAAR Particulate Matter Supersites Program and Related Studies, February 7-11, 2005, Atlanta, Georgia. Abstract 13C-4, pp. 113.

Rong, Y. (2000) Statistical methods and pitfalls in environmental data analysis. Environmental Forensics 1, 213–220.

Sammon Jr., J.W. (1969) A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, C-18, 401–409.

San José, R., Peréz, J.L., Morant, J.L., González, R.M. (2009) Modeling of Pollutants in Atmospheric Environmental Systems. In Hanharan, G. (Ed.), Modeling of Pollutants in Complex Environmental Systems. Hertfordshire: ILM Publications.

Sarle, W.S. (1995) Stopped training and other remedies for overfitting. Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, pp. 352–360.

Seppelt, R. (2003) Computer-based Environmental Management. VCH-Wiley: Wenheim, Germany.

Schafer, J.L. (1997) Analysis of incomplete multivariate data. Monographs on Statistics and Applied Probability No. 72., Chapman & Hall, London.

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterto, T., Vondracek, Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., Doyle, M., (2003) A rigorous inter-comparison of ground-level ozono predictions. Atmospheric Environment 37, 3237–3253.

Schneider, T. (2001) Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate 14, 853–871.

Schultz, T.W., Cronin, M.T.D., Walker, J.D., Aptula, A.O. (2003) Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. Journal of Molecular Structure (Theocem) 622, 1–22.

Simula, O., Vesanto, J., Alhoniemi, E. and Hollmèn, J. (1999) Analysis and Modeling of Complex Systems using the Self-Organizing Map. In Kasabov,N. and Kozma, R. (Eds.), Neuro-Fuzzy Techniques for Intelligent Information Systems. Springer Verlag.

Siwek, K., Osowski, S. and Sowinski, M. (2010). Neural predictor ensemble for accurate forecasting of PM10 pollution. Proceedings of the 2010 International Joint Conference on Neural Network (IJNN).

Smola, A. J. and Schölkopf, B. (2003) A tutorial on support vector regression, NeuroCOLT2 Technical Report NC2-TR-1998-030.

Snee, R.D. (1977) Validation of regression models: Methods and examples. Technometrics 19, 415–428.

Solomatine, D.P. (2005) Data-driven modelling and computational intelligence methods in hydrology. In M. Anderson (Ed.), Encyclopedia of hydrological sciences. New York: Wiley.

Solomatine, D.P. and Ostfeld, A. (2008) Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics 10, 3–22.

Srinivas, N. and Deb, K. (1994). Multiobjective optimization using nondominated sorting in genetic algorithms. Evolutionary Computation 2, 221–248.

Sulkava, M. (2008) Learning from environmental data: methods for analysis of forest nutrition time series. D.Sc. Thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D24, Espoo.

Thomsen, M. (2001) QSARs in Environmental Risk Assessment – Interpretation and validation of SAR/QSAR based multivariate data analysis. The doctoral dissertation. Department of Environmental Chemistry, National Environmental Research Institute.

Tichý, M. and Rucki, M. (2009) Validation of QSAR models for legislative purposes. Interdisciplinary Toxicology 2(3), 184-186.

Turias, I.J., González, F.J., Martin, M.L. and Galindo, P.L. (2007) Prediction models of CO, SPM and SO2 concentrations in the Campo de Gibraltar Region, Spain: a multiple comparison strategy. Environmental Monitoring and Assessment 143, 131-146.

Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer, New York.

Vauhkonen, J. (2010) Estimating single-tree attributes by airborne laser scanning: methods based on computation geometry of the 3-D point data. Dissertations Forestales 104.

Veltheim, T. (1987) Pituusmallit männylle, kuuselle ja koivulle, M.S. thesis, Department of Forest Mensuration and Management, University of Helsinki, Helsinki, Finland.

Voss, R.F. (1991), In Peitgen, H.-O. and Saupe, D., (Eds.), The Science of Fractal Images, Springer-Verlag, New York, 1991.

Wehr, A. and Lohr, U. (1999) Airborne laser scanning – an introduction and overview. ISPRS Journal of Photogrammetry and Remote Sensing 54, 68–82.

Weisberg, S. (1985) Applied Linear Regression, 2nd ed., John Wiley, New York.

Willmott, C. J. (1981) On the validation of models. Physical Geography 2, 184–194.

Willmott, C. J., Ackleson, S., Davis, R., Feddema, J., Klink, K., Legates, D., O'Donnell, J. and Rowe, C. (1985) Statistics for the evaluation and comparison of models. Journal of Geophysical Research 90 (C5), 8995–9005.

Wold, S., Esbensen, K. and Geladi, P. (1987) Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2, 37–52.

Worth, A.P., Bassan, A., De Bruijn, J., Gallegos Saliner, A., Netzeva, T., Patlewicz, G. Pavan, M., Tsakovska, I. and Eisenreich, S. (2007) The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. SAR and QSAR in Environmental Research 18(1), 111–125.

Yao, X. (1999) Evolving artificial neural networks. Proceedings of the IEEE Transactions on Neural Networks 87 (9), 1423–1447.

Yu, X. and Liong, S.-Y. (2007) Forecasting of hydrologic time series with ridge regression in feature space. Journal of Hydrology 332, 290–302.

Åström, K and K Wittenmark, K (1990). Computer Controlled Systems: Theory and Design. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

**Harri Niska**

# *Predictive Data-Driven Modeling Approaches in Environmental Management Decision-Making*

Nowadays, there is an increasing need for powerful and reliable computational models that can be used to support decision-makers in managing and regulating environmental issues. This work provides novel data-driven modeling approaches, which rely mainly on the methods of computational intelligence, for solving complex prediction problems associated with urban air quality control, chemical risk assessment, and forest inventory. It is shown that the computational approaches studied entail many inherent benefits for environmental data processing and modeling, providing thus potential alternatives to the conventional procedures used in environmental management decision-making.

UNIVERSITY OF
EASTERN FINLAND

Publications of the University of Eastern Finland
*Dissertations in Forestry and Natural Sciences*

ISBN 978-952-61-0646-5