Helsinki University of Technology
Dissertations in Computer and Information Science
Espoo 2008 Report D24

LEARNING FROM ENVIRONMENTAL DATA:

METHODS FOR ANALYSIS OF FOREST NUTRITION TIME SERIES

Mika Sulkava

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 18th of January, 2008, at 12 o'clock noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Distribution:
Helsinki University of Technology
Department of Information and Computer Science
P.O.Box 5400
FI-02015 TKK
FINLAND
Tel. +358 9 451 3272
Fax +358 9 451 3277
http://www.cis.hut.fi

Available in PDF format at
http://lib.hut.fi/Diss/2008/isbn9789512291540/

ISBN 978-951-22-9153-3 (printed version)
ISBN 978-951-22-9154-0 (electronic version)
ISSN 1459-7020

Helsinki University Print
Helsinki 2008

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

Keywords: data analysis, data mining, time series, forest, foliage, nutrient, environmental informatics, environmental statistics, environmental monitoring, clustering, self-organizing map, sparse regression, weighted regression.

ABSTRACT

Data analysis methods play an important role in increasing our knowledge of the environment as the amount of data measured from the environment increases. This thesis fits under the scope of environmental informatics and environmental statistics. They are fields, in which data analysis methods are developed and applied for the analysis of environmental data.

The environmental data studied in this thesis are time series of nutrient concentration measurements of pine and spruce needles. In addition, there are data of laboratory quality and related environmental factors, such as the weather and atmospheric depositions.

The most important methods used for the analysis of the data are based on the self-organizing map and linear regression models. First, a new clustering algorithm of the self-organizing map is proposed. It is found to provide better results than two other methods for clustering of the self-organizing map. The algorithm is used to divide the nutrient concentration data into clusters, and the result is evaluated by environmental scientists. Based on the clustering, the temporal development of the forest nutrition is modeled and the effect of nitrogen and sulfur deposition on the foliar mineral composition is assessed.

Second, regression models are used for studying how much environmental factors and properties of the needles affect the changes in the nutrient concentrations of the needles between their first and second year of existence. The aim is to build understandable models with good prediction capabilities. Sparse regression models are found to outperform more traditional regression models in this task.

Third, fusion of laboratory quality data from different sources is performed to estimate the precisions of the analytical methods. Weighted regression models are used to quantify how much the precision of observations can affect the time needed to detect a trend in environmental time series. The results of power analysis show that improving the quality may decrease the time needed for detection of the trend by many years.

The data analysis methods developed and applied in this thesis are found to produce results which are understandable for the environmental scientists. They are, therefore, useful for studying the condition of the environment and evaluating the possible causes for changes in it.

Sulkava, M. (2008): **Ympäristödatasta oppiminen: menetelmiä metsän ravintoaikasarjojen analyysiin**. Tohtorin väitöskirja, Teknillinen korkeakoulu, Dissertations in Computer and Information Science, raportti D24, Espoo, Suomi.

Avainsanat: data-analyysi, tiedonlouhinta, aikasarja, metsä, neulasto, ravinne, ympäristöinformatiikka, ympäristötilastotiede, ympäristönseuranta, ryvästys, itseorganisoiva kartta, harva regressio, painotettu regressio.

TIIVISTELMÄ

Data-analyysimenetelmät ovat tärkeässä osassa ympäristöä koskevan tiedon kartuttamisessa, kun ympäristöstä mitatun datan määrä kasvaa. Tämä väitöskirja kuuluu ympäristöinformatiikan ja ympäristötilastotieteen aloihin. Näillä tieteenaloilla data-analyysimenetelmiä kehitetään ja sovelletaan ympäristödatan analysointiin.

Tässä väitöskirjassa tutkittu ympäristödata on aikasarjoja männyn- ja kuusenneulasten ravinnepitoisuusmittauksista. Lisäksi väitöskirjassa on käytetty laboratorioiden laadusta sekä aiheeseen liittyvistä ympäristötekijöistä kuten säästä ja laskeumista mitattua dataa.

Tärkeimmät datan analysoinnissa käytetyt menetelmät perustuvat itseorganisoivaan karttaan ja lineaarisiin regressiomalleihin. Ensiksi esitellään uusi itseorganisoivan kartan ryvästysalgoritmi. Sen havaitaan tuottavan parempia tuloksia kuin kaksi muuta itseorganisoivan kartan ryvästysmenelmää. Algoritmia käytetään jakamaan ravinnepitoisuusdata ryppäisiin. Tämän jälkeen ympäristötieteilijät arvioivat tulosta. Ryvästyksen pohjalta mallitetaan metsien ravinteiden ajallista kehitystä ja arvioidaan typpi- ja rikkilaskeuman vaikutusta neulaston kivennäiskoostumukseen.

Toiseksi regressiomalleja käytetään tutkimaan, kuinka paljon ympäristötekijät ja neulasten ominaisuudet vaikuttavat muutoksiin neulasten ravinnepitoisuuksissa niiden ensimmäisen ja toisen olemassaolovuoden välillä. Tavoitteena on rakentaa ymmärrettäviä malleja, joiden ennustuskyky on hyvä. Harvojen regressiomallien todetaan suoriutuvan tästä tehtävästä perinteisempiä regressiomalleja paremmin.

Kolmanneksi eri lähteistä peräisin olevaa laboratorioiden laatua mittaavaa dataa yhdistetään, ja sen avulla lasketaan analyysimenetelmien tarkkuudet. Painotettuja regressiomalleja käytetään määrittämään, kuinka paljon laboratorioiden laatu voi vaikuttaa trendin havaitsemiseen ympäristöaikasarjoista. Voima-analyysin tulokset osoittavat, että laadun parantaminen voi vähentää havaitsemiseen tarvittavaa aikaa useilla vuosilla.

Tässä väitöskirjassa kehitettyjen ja käytettyjen data-analyysimenetelmien todetaan tuottavan tuloksia, jotka ovat ymmärrettäviä ympäristötieteilijöille. Ne ovat siksi hyödyllisiä tutkittaessa ympäristön kuntoa ja arvioitaessa sen muutosten mahdollisia syitä.

Preface

The work leading to this thesis has been done in the Laboratory of Computer and Information Science of Helsinki University of Technology (TKK). The work in the laboratory has been funded by the Adaptive Informatics Research Centre and earlier by the Neural Networks Research Centre. The work has also been partly funded by the Graduate School of Department of Computer Science and Engineering, and the Academy of Finland in the research project: Analysis of dependencies in environmental time-series data (AD/ED). I am thankful to the above-mentioned organizations for financing my work. During my graduate studies I have also had the opportunity to participate the courses and other events organized by Helsinki Graduate School in Computer Science and Engineering (Hecse). In addition, I want to show my gratitude to Emil Aaltonen foundation, Foundation of Technology (TES), and the Finnish Foundation for Economic and Technology Sciences (KAUTE) for the personal grants.

I wish to thank my instructor Dr. Jaakko Hollmén and supervisor Prof. Olli Simula for their encouragement and support during my graduate studies and writing of this thesis. I also wish to thank Academician Teuvo Kohonen, Prof. Heikki Mannila, and the head of the laboratory Prof. Erkki Oja for their pioneering work and for making it possible for me to do research in the laboratory.

I want to thank also other people from the laboratory for cooperation. I would like to thank the co-authors of the publications of this thesis Dr. Juha Vesanto and Jarkko Tikka and my other closest co-workers Pasi Lehtimäki, Dr. Timo Similä, Hannes Heikinheimo, and Mikko Korpela for fruitful collaboration and for the inspiring company at the laboratory.

A large part of the work related to this thesis has been done in collaboration with the Finnish Forest Research Institute (Metla). I am thankful to Dr. Sebastiaan Luyssaert—currently in University of Antwerp—for initiating this line of research and for smooth collaboration. I also thank Prof. Hannu Raitio, Dr. Pasi Rautio, and Dr. Päivi Merilä from Metla, Prof. Ivan A. Janssens from University of Antwerp, and Alfred Fürst from the Austrian Federal Office and Research Centre for Forests (BFW) for collaboration.

I have received a lot of constructive feedback on the manuscripts of this thesis. I am thankful to Prof. Olli Simula, Dr. Jaakko Hollmén, Pasi Lehtimäki, Jarkko Tikka, and Pauliina Pesola for proofreading this thesis. I also thank the pre-examiners Prof. Sašo Džeroski and Dr. Alfredo Vellido for reviewing this thesis and for their positive and encouraging comments. In addition, I wish to thank Prof. Thomas Martinetz for accepting to act as the opponent at the defense.

Finally, I want to thank my parents Sauli and Raili, my relatives and friends, and of course my beloved Pauliina for being so supportive of my studies, work, and writing of this thesis. Many thanks; no man does it all by himself.

Maunula, December 19, 2007

Mika Sulkava

Abbreviations and notations

In this thesis, the following abbreviations and notations are used. Scalar values are denoted by lower case symbols (e.g., a), vectors by lower case boldface symbols (e.g., $\mathbf{m}$), and matrices by upper case boldface letters (e.g, $\mathbf{A}$). An estimate of a parameter or prediction of a value is denoted by the hat (e.g., $\hat{y}$).

ICP Forests	International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests
IUFRO	International Union of Forest Research Organizations
KDD	knowledge discovery in databases
LARS	least angle regression
lasso	least absolute shrinkage and selection operator
Metla	Finnish Forest Research Institute
OLS	ordinary least squares
SOM	self-organizing map
U-matrix	unified distance matrix
Al	aluminum
β	regression coefficient
C	foliar age class of current year needles
$C + i$	foliar age class of i-year-old needles
Ca	calcium
d	number of variables
ϵ	error term
i, j, q	indices
$\mathbf{m}$	prototype vector of the SOM
n	number of observations
N	nitrogen
$P(A\|B)$	conditional probability of A given B
s	state of a Markov chain
S	sulfur
σ	standard deviation
t	time index
$\mathbf{W}$	weight matrix of a weighted regression model
x	observed value of data; regressor variable in case of regression
y	observed value of response variable
$\|\| \cdot \|\|$	Euclidean norm

Contents

Chapter 1

Introduction

1.1 Scope of the thesis

The environment is changing in many parts of the earth. Some of these changes are caused by human activity. The changes can have great influence on people's lives and also on other forms of life; see, e.g., Emanuel (2005). Human-induced or not, it is important to identify the causes and possible consequences of the changes in order to be able to act in such a way that the negative effects of these changes can be reduced.

During the last decades computers have been developed at a fast rate and new data analysis methods and modeling techniques have been proposed (Hand et al., 2001). Also, data collection has become less laborious, which has increased the amount of available environmental data. This development, i.e., better analyses with more data using faster computers, has made it possible to gain a better understanding on the environment. It is not, however, a system whose functioning is already completely understood. In addition, the analysis tools still need to be developed for new problems.

This thesis considers environmental informatics, which is a new discipline between information technology and environmental sciences. Some of the contents of the thesis also fit under the scope of environmental statistics or environmetrics. A central part of both disciplines is applying and developing data analysis methods for environmental problems. This is also the main theme of this thesis.

The research problem of this work is to study how data analysis methods can be efficiently used to extract relevant new information from multidimensional environmental time series data. Forest nutrition data from an environmental monitoring program is studied in different ways. The data consists of measurements of coniferous foliar samples, i.e., samples of tree needles. Also, other complemen-

tary data sets containing information on weather, atmospheric deposition, and laboratory quality are studied.

The aim of the research is to develop and apply data analysis methods to different kinds of environmental problems, which are related to nutritional status of forest foliage. The methods are chosen and designed to meet various special needs of the problems. The most important research methods include the self-organizing map (SOM), clustering, regression models, trend tests, power analysis, and cross-validation.

1.2 Publications of the thesis

This thesis consists of an introductory part and the following seven peer-reviewed original publications.

1. Juha Vesanto and Mika Sulkava (2002). Distance matrix based clustering of the self-organizing map. In Dorronsoro, J. R., editor, *Artificial Neural Networks – ICANN 2002: International Conference, Proceedings*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain. Springer-Verlag.

2. Mika Sulkava and Jaakko Hollmén (2003). Finding profiles of forest nutrition by clustering of the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 243–248, Kitakyushu, Japan.

3. Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén (2004). Evaluation of forest nutrition based on large-scale foliar surveys: are nutrition profiles the way of the future? *Journal of Environmental Monitoring*, 6(2):160–167.

4. Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén (2005). Are N and S deposition altering the chemical composition of Norway spruce and Scots pine needles in Finland? *Environmental Pollution*, 138(1):5–17.

5. Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén (2006). Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modelling*, 191(1):118–130.

6. Mika Sulkava, Pasi Rautio, and Jaakko Hollmén (2005). Combining measurement quality into monitoring trends in foliar nutrient concentrations. In Duch, W., Kacprzyk, J., Oja, E., and Zadrożny, S., editors, *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005: 15th International Conference, Proceedings, Part II*, volume 3697 of *Lecture Notes in Computer Science*, pages 761–767, Warsaw, Poland. Springer-Verlag.

7. Mika Sulkava, Sebastiaan Luyssaert, Pasi Rautio, Ivan A. Janssens, and Jaakko Hollmén (2007). Modeling the effects of varying data quality on trend detection in environmental monitoring. *Ecological Informatics*, 2(2):167–176.

In the text, these publications are referred to using the numbering above, e.g., Publication 7.

1.3 The author's contributions

The most important scientific contributions of the author presented in this thesis are the following:

- An algorithm is developed for automatic clustering of the self-organizing map. The algorithm is found to produce better results than two other approaches for clustering of the self-organizing map.
- The clustering algorithm is used to find ecologically relevant clusters in forest nutrition data. Condition and development of the forests is analyzed based on the clusters. The development of forests is characterized using transition matrices and graphs.
- Temporal changes of the element concentrations in needles is analyzed. Linear sparse regression models are used for finding the most important environmental factors and properties of needles, which affect the aging of the needles. The models are found to be better suited for the task than other simpler linear regression approaches.
- The time needed to detect trends in heteroscedastic time series data is quantified using power analysis and weighted regression models. Laboratory quality data from multiple sources is used to estimate the precisions of the chemical analysis methods of the laboratories analyzing the forest nutrition data. Based on this information, weighted regression models are used to assess how well statistically significant trends can be detected in the observed forest nutrition data.

Publication 1 presents an algorithm for automatic clustering of the self-organizing map. The author participated in designing the algorithm with Vesanto. The experiments were designed jointly with Vesanto and conducted for the most part by the author. The paper was written for the most part by Vesanto.

The author developed mostly the idea of using clustering of the self-organizing map to find representative group nutrition profiles in the forest nutrition data in Publications 2, 3, and 4. In addition, the author developed the idea to analyze the temporal behavior of the forests with transition matrices (Publications 2 and

3) and graphs (Publication 2) showing probabilities of transitions between the clusters. The author designed most of the experiments for Publications 2 and 3 and conducted all of them. The author and Luyssaert designed and conducted most of the experiments for Publication 4. Publication 2 was written by the author, Publication 3 jointly with Luyssaert, and Publication 4 mostly by Luyssaert. Luyssaert and Raitio assessed the environmental implications of the results in Publications 3 and 4.

In Publication 5 the author developed mostly the idea of using sparse regression models to find the most important environmental factors affecting the aging of tree needles. The author designed the experiments jointly with Tikka and Hollmén, conducted parts of them, and wrote most of the paper.

In Publications 6 and 7 the idea of quantifying the time needed for trend detection in time series with varying data quality was designed jointly with Rautio, Luyssaert, and Hollmén. The author designed most of the experiments, conducted all of them, and wrote most of both publications.

1.4 Structure of the thesis

The following chapters of the introductory part are organized as follows. The ecological background of the work and the data are presented in Chapter 2. The methodological background is presented in Chapter 3. The analyses are explained in Chapter 4. Finally, Chapter 5 summarizes the work and concludes the introductory part. In the printed version the publications are attached after the introductory part.

Chapter 2

Environment as a source of time series data

The environment provides all life forms, including the human, with the conditions of life. The environment can also be considered important by itself. Mankind causes changes to the environment, and extreme changes may cause serious problems. Therefore, it is important to analyze and understand how the environment works, what causes changes in it, and what are their consequences. Environmental sciences study these kinds of issues.

2.1 Characteristics of environmental data

The environment is a highly complex system. Even the smallest ecosystems have many sources of complexity. Green and Klomp (1999) mention the following:

- The spatial scale
- The temporal scale
- The number of organisms
- Criticality
- Non-linear interactions and feedback loops
- Human influence on natural systems

Criticality above can be clarified by an example of phase changes. If small patches are removed randomly from a forest and if the clearing continues, at some critical point the forest breaks into isolated fragments.

Environmental data sets are often multivariate time series (Günther, 1998), which are also space dependent (Page and Rautenstrauch, 2001). Considerable amount of data is in analog form (Günther, 1998), and if these data are needed for analysis, acquiring them in digital format is somewhat laborious.

Metadata is data about data, i.e., it describes information processing and supply. It may tell what has been measured and how, where it is available, and how it can be technically accessed (Page, 1995). This information may increase the longevity of the data set, facilitate reuse and sharing of the data with others, and expand the temporal scale of ecological inquiry (Michener, 2000a). Metadata can also contain information about the uncertainty of the data. Additionally, data fusion, i.e., combining data from multiple sources, can be used to assess the uncertainty of the data (Günther, 1998). Data fusion, however, can be difficult in some cases due to, e.g., nonstandardized measurements (Slagle, 1994).

Continuous evaluation of environmental data streams is called environmental monitoring (Günther, 1998). When the environmental monitoring data is stored year after year, the size of the data sets often becomes large.

2.2 Forest ecosystems

Forests are complex ecosystems, which in addition to having a high density of trees, affect the earth's water cycle and properties of the soil, store carbon, and act as important animal habitats. Large data sets from long-term monitoring programs are required in order to gain an insight into the condition of forests and in order to assess the future development of forests under the present and predicted environmental scenarios. Currently there are several large-scale forest monitoring programs, e.g., United Nations Economic Commission for Europe (UNECE) International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests, 2007) in Europe and North America, Forest Focus (2007) in the European Union, Acid Deposition Monitoring Network in East Asia (EANET, 2007), and United States Department of Agriculture (USDA) Forest Service – National Forest Health Monitoring Program (FHM, 2007). Due to the relationships between the environment and the foliar mineral composition, these programs monitor, among other ecosystem components, the mineral composition of tree foliage. As a result of the monitoring, the programs have over the years built up large data sets of the mineral composition of tree foliage. Large-scale monitoring requires effective information management. For an example, Liff et al. (1994) describe the information management system, which is used for handling the data in FHM.

ICP Forests monitors the properties of the forests in the following surveys: crown condition, foliar chemistry, soil chemistry, tree growth, ground vegetation, stand structure including deadwood, epiphytic lichens, soil solution chemistry, atmospheric deposition, ambient air quality, meteorology, phenology, litterfall, and

remote sensing. This thesis contains analyses of forest nutrition data of the ICP Forests. The data was provided by the Finnish Forest Research Institute (Metla, 2007).

2.3 Forest nutrition

Controlled experiments have shown that the mineral composition of tree foliage is related to many environmental factors including nitrogen (N) and sulfur (S) deposition, climatic variation, drought, and the ozone (O_3) and carbon dioxide (CO_2) concentrations. In a long-term monitoring program, the observations are made in complex real-world conditions instead of controlled conditions of designed experiments. Therefore, making conclusions based on monitoring data is more challenging.

The forest nutrition data used in this study is from a foliar survey that was initiated as a part of the ICP Forests program. Needle samples were collected from 38 stands of the Finnish Level I network of The Finnish Forest Condition Monitoring Programme (FFCMP, 2007) (Forest Focus / ICP Forests). Seventeen of the stands were dominated by Norway spruce (*Picea abies* (L.) Karst.) and 21 by Scots pine (*Pinus sylvestris* L.). The age of the trees in the stands was between 30 and 342 years. The stands were located in different parts of Finland and they were sampled annually between 1987 and 2003. 29 % of the maximum possible number of 646 composite samples (38 stands $\times$ 17 years) were missing. The missing values were especially concentrated in years 1990 and 1991. In those years, 84 % and 61 % of the composite samples were missing, respectively. Foliar concentrations of 12 elements expressed as percent by mass (or mg/g or μg/g) and the average mass per needle (or per 1000 needles) were determined on composite samples of 10 trees per stand. The elements are: aluminum (Al), boron (B), calcium (Ca), copper (Cu), iron (Fe), potassium (K), magnesium (Mg), manganese (Mn), nitrogen (N), phosphorus (P), sulfur (S), and zinc (Zn). In Publications 2, 3, 5, and 4, data was available for years 1987–2000, in Publication 6 for years 1987–2002, and in Publication 7 for years 1987–2003. Sample collection, pre-treatment, and chemical analysis are presented in detail in Publication 3.

The growth of the trees is controlled by foliar element concentrations a_i, $i = 1, \ldots, d-1$ and contents $a_i M$, where M is the foliar mass and d is the number of variables, i.e., element concentrations and foliar mass. In the forest nutrition data described earlier, the number of variables $d = 13$. In Publication 3, the nutrition profile of a tree or stand is defined as the nutrient status, which accounts for all element concentrations, contents, and interactions between two or more elements. The nutrition profile vector of deciduous trees, i.e., trees which lose their leaves

seasonally, is as follows:

$$\mathbf{p} = \begin{bmatrix} a_1 \\ \vdots \\ a_{d-1} \\ M \end{bmatrix}. \tag{2.1}$$

Evergreen trees have a number $L + 1$ of different foliar age classes: C, $C + 1$, ..., $C + L$. Here C denotes the leaves or needles, which were grown in the current year, $C+1$ denotes the leaves or needles, which were grown in the previous year, etc. The nutrition profile of evergreen trees is therefore more complicated:

$$\mathbf{p} = \begin{bmatrix} a_{1,C} \\ \vdots \\ a_{d-1,C} \\ M_C \\ \vdots \\ a_{1,C+L} \\ \vdots \\ a_{d-1,C+L} \\ M_{C+L} \end{bmatrix}. \tag{2.2}$$

Figure 2.1 visualizes the structure of the forest nutrition data. A row in the two adjacent matrices is a nutrition profile of foliar age classes C and $C+1$, which were determined in the foliar surveys. Figure 2.2 shows the average S concentration of all stands for the 17 years. Each point in the figure represents the average of the 38, or less, measurements made in the corresponding year. A decreasing trend in average S concentration can be seen in the figure. There is also other variation between the years.

2.4 Laboratory quality

Laboratories of Metla analyzed the foliar samples of all the years. As part of an integrated quality assurance system the laboratories carried out repeated measurements of certified reference samples and participated in international interlaboratory tests arranged by International Union of Forest Research Organizations (IUFRO, 2007) and ICP Forests. The quality of the instrumental analysis of the laboratories of Metla has improved partly due to the quality assurance system.

The repeated measurements of reference samples and in the interlaboratory tests allow to estimate the accuracy and precision of the analytical methods as applied by the laboratories. Accuracy denotes the systematic component of the error in the measurements and precision denotes the random component. In other words, accuracy is the bias between the true value and the mean of the observed values

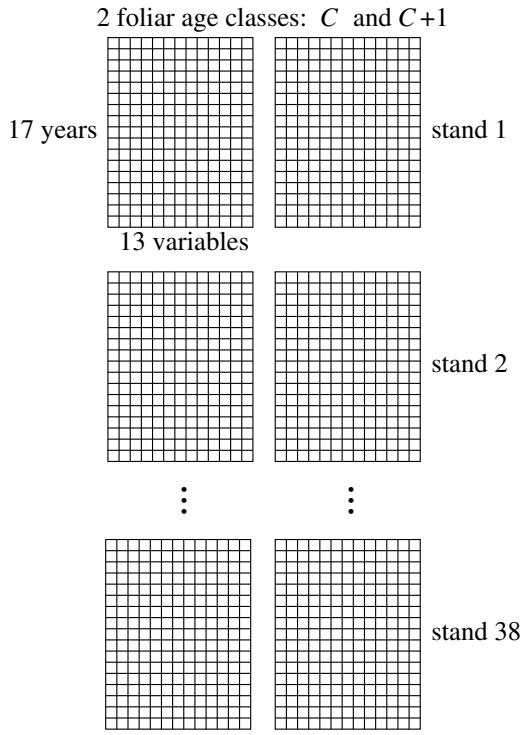


Figure 2.1: Structure of the forest nutrition data. 13 variables have been measured annually from needles of two foliar age classes in 38 stands during 17 years.

and precision is the variation of the observed values. For more details on accuracy, precision, and laboratory quality, see, e.g., Berthouex and Brown (2002). The quality of measurements in terms of accuracy and precision is illustrated in Figure 2.3 with a target analogy and with a distribution of repeated measurements.

The accuracy is estimated as the average deviation of the concentrations determined by laboratories of Metla and either the known correct concentrations or the average concentrations of the same samples determined by all laboratories which participated in the interlaboratory test and reported acceptable results. The precision is calculated as the unbiased estimate for the standard deviation of the concentrations determined by the laboratories of Metla.

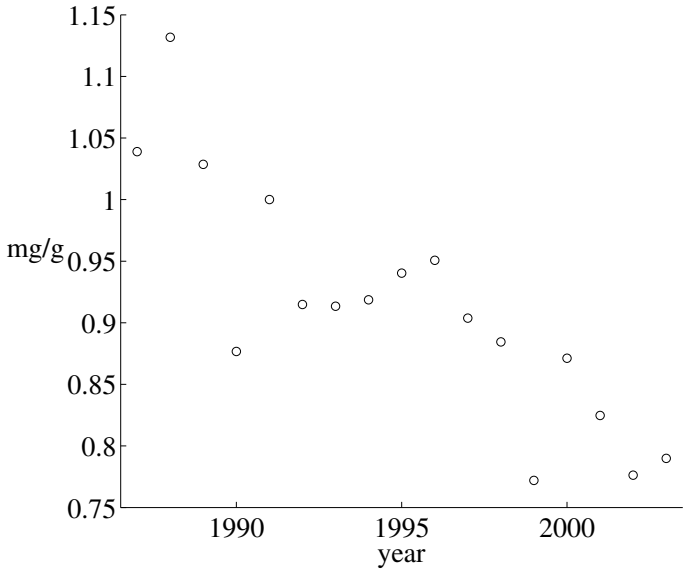


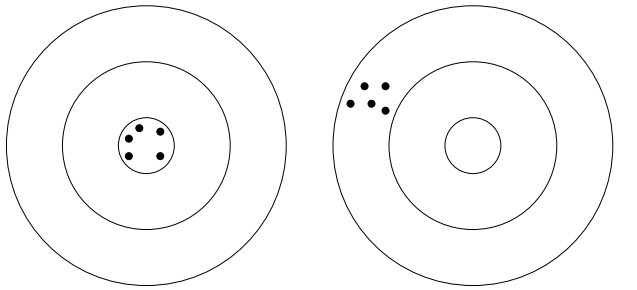
Figure 2.2: Average S concentration in all stands. The high number of missing values may have affected the average values of years 1990 and 1991.

2.5 Other environmental factors

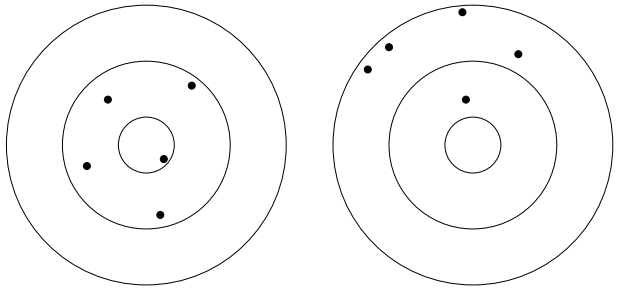
Many environmental factors are interconnected. For example, weather and anthropogenic (human-induced) effects such as atmospheric deposition, affect the growth and condition of forests. On the other hand, forests take up carbon from the atmosphere as they grow (Magnani et al., 2007), which again may affect the climate (Schulze et al., 2000).

In this study, there was also other environmental data available. First, there was weather data. Monthly average temperatures and precipitation sums were obtained from the Finnish Meteorological Institute (FMI, 2007) for weather stations located near the stands. This data covered years 1971–2000 and it was rather complete: 0.6 % of the monthly average temperatures and 3 % of monthly precipitation sums were missing.

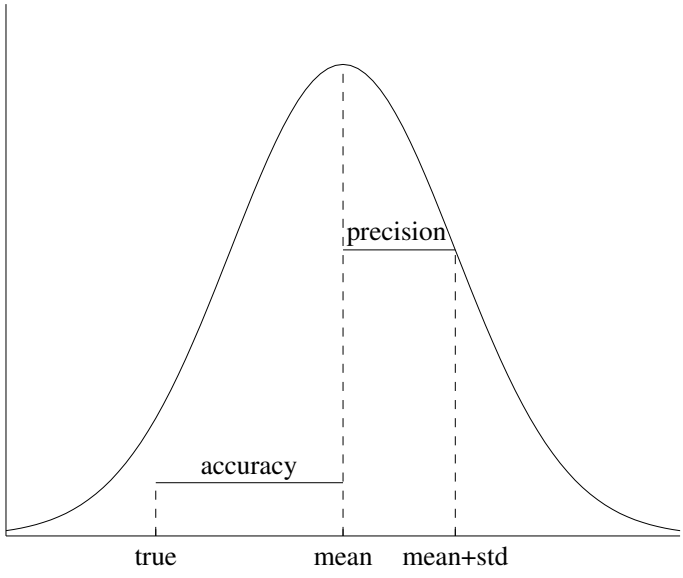
Second, there was deposition data of the Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe (EMEP, 2007). Modeled yearly total deposition values of oxidized and reduced N and oxidized S were available for years 1985–1996 on a $150 \times 150\,\mathrm{km}^2$ grid. The deposition values in the stands were obtained using interpolation (Publication 4).



(a) Accurate and precise (b) Not accurate but precise



(c) Accurate but not precise (d) Not accurate and not precise



(e) Distribution of repeated measurements

Figure 2.3: Accuracy and precision visualized with a target analogy (a–d) and distribution of repeated measurements (e). Accuracy is the bias between the true value and the mean of the distribution of repeated measurements. Precision measures the width of the distribution (here standard deviation, std).

Chapter 3

Fields of research and methods of environmental data analysis

This chapter concerns the analysis of environmental data. In the first section, fields of research related to learning and drawing conclusions based on observed data are reviewed. In the second section, fields which focus on environmental data analysis are reviewed. In the third section, the most important data analysis methods used in this thesis are presented.

3.1 Fields related to learning from data

3.1.1 General frameworks of learning

Learning from data is an important part of statistics, data mining, and artificial intelligence (Hastie et al., 2001). A part of artificial intelligence, machine learning, is about developing techniques which enable computers to learn from examples. Machine learning can also, less ambitiously, be seen as using modern methods to learn from data (Cherkassky and Mulier, 1998). The latter definition suits better the scope of this thesis. Learning can be either supervised or unsupervised, meaning that either there is some a priori information available about the true outputs or not. For various theories of learning, see Phillips and Soltis (2004).

Knowledge discovery in databases (KDD) is defined by Frawley et al. (1991) as:

> *the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.*

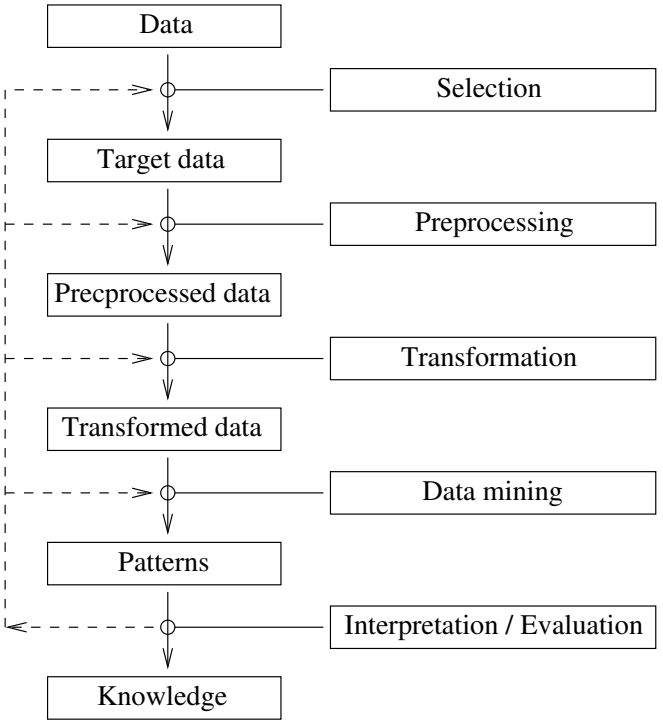


Figure 3.1: The steps of the KDD process. The figure is adapted from Fayyad et al. (1996).

More generally, KDD can be seen to refer to (Fayyad et al., 1996):

the overall process of discovering useful knowledge from data.

The steps of the KDD process are shown in Figure 3.1. It is an interactive and iterative process. Before taking the first step of selecting the target data set, some prior knowledge of the application domain should be obtained. Next, the data are preprocessed by, e.g., removing noise and handling missing data. In the transformation step, data reduction and projection methods are used to reduce the number of variables and to find useful features for the task. In the fourth step, data mining methods are selected and used to find patterns in the data, which match the goals of the KDD process. The last step is the interpretation of the patterns involving, e.g., visualization of the patterns and the data. The next step is acting on the discovered knowledge or if the results are not satisfactory, returning to one of the previous steps to make some better choices.

In addition to the KDD process described above, there are other similar models, e.g., the Cross Industry Standard Process for Data Mining – CRISP-DM (Chapman et al., 2000) and ladder of inference (Argyris, 1982; Senge et al., 1996), which

describe the process of acquiring knowledge or taking actions based on observed data.

3.1.2 Statistics and hypothesis testing

Data analysis in statistics is often confirmatory, meaning that hypotheses about nature are tested and either confirmed or rejected. In addition, data is usually collected from carefully designed experiments, which involve sampling, randomization, replication, controlling for confounding variables, etc. The general experimental procedure of the scientific method contains the following steps (Dowdy et al., 2004):

1. State the problem.
2. Formulate the hypothesis.
3. Design the experiment or survey.
4. Make observations.
5. Interpret the data.
6. Draw conclusions.

The elements of statistical learning are shown in Figure 3.2. Knowledge about nature increases as the phases are iterated: new hypotheses are defined, new experiments are designed, and new data is collected until the problem is solved. This iterative model of learning—similar to the KDD process described earlier—may represent reality better than the basic linear scientific method.

Statistical hypothesis testing includes two complementary hypotheses: the null hypothesis and the alternative hypothesis. Often the null hypothesis is that there is no effect and the alternative hypothesis is the complement, i.e., that there is any kind of effect (Hand et al., 2001). The number of statistical tests designed for testing various hypotheses is very high. Commonly used statistical tests are, e.g., t-tests, F-tests (Jørgensen, 1993), and permutation tests (Good, 2005).

3.1.3 Data mining

In the past decades, data collection has become easier and the amount of data has increased. The data, however, is not always from designed experiments, and in many cases formulating well defined hypotheses is difficult. Thus, the general experimental procedure cannot be followed as such.

If the situation is viewed in the framework of the elements of learning in Figure 3.2, the problem is probably not completely defined, there is no clear hypothesis, and

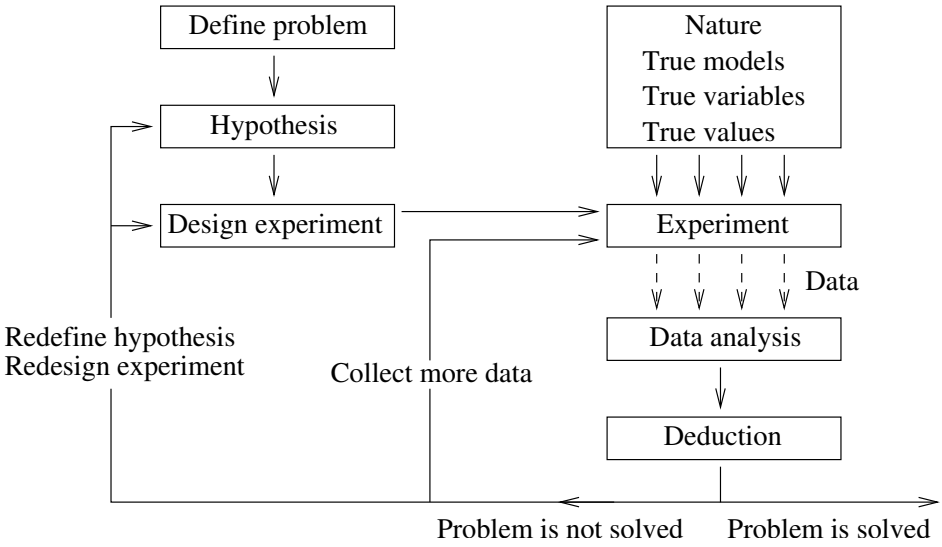


Figure 3.2: Elements of learning. The figure is adapted from Berthouex and Brown (2002).

no designed experiment. In this case, the first step is data analysis, which after deduction could lead to define the problem and hypothesis and probably also to design an experiment and collect more data. In this kind of situation, the first step of analyzing the data may be called data mining. In particular, data mining involves retrospective analyses of data (Glymour et al., 1997). Hand et al. (2001) define 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.*

Also, Glymour et al. (1997) state that understandability is often more important in data mining than accuracy of predictability. More specifically, the tasks of data mining are (Hand et al., 2001):

1. Exploratory data analysis
2. Descriptive modeling
3. Predictive modeling: classification and regression
4. Discovering patterns and rules
5. Retrieval by content

Thus, data mining is an interdisciplinary field where statistics plays an important role. However, instead of hypothesis testing, data mining is more about generating the hypotheses. The contents of this thesis—especially Publications 2, 3, 4, and 5, and also partly Publications 6 and 7—fall into the four first categories above. Some of these tasks are explained in more detail later in this chapter.

3.1.4 Exploratory data analysis

As noted earlier, not all data analysis tasks are confirmatory. Tukey (1980) argues that we should

> *think about science and engineering more broadly than the narrow, inadequate paradigm of a straight line from question to answer.*

The goal of exploratory data analysis (Tukey, 1977), then, is to explore the data without exact knowledge of what we are looking for. Exploratory data analysis is defined by Hand et al. (2001) as:

> *data-driven hypothesis generation.*

In this task, various visualization methods (Keim, 2002) play a key role. For many people just looking at the numeric data does not help much in hypothesis generation. A good visualization shows the data to the analyst in a way, which allows effective hypothesis generation. Additionally, dimensionality reduction (Cumming and Wooff, 2007) and projection methods (Similä, 2007) are useful tools in exploratory data analysis. They project multidimensional data into a lower dimension and thus, make visualization and exploration of the data easier. For various tools of exploratory data analysis, see, e.g., Hoaglin et al. (2000).

3.1.5 Pattern recognition

Pattern recognition is a part of machine learning, which aims to discover automatically regularities in data using computer algorithms (Bishop, 2006). These regularities can be used to classify the data into categories or classes using features of the objects (Duda et al., 2001; Theodoridis and Koutroumbas, 1999) or to find some regression function, which instead of the discrete class information, explains a continuous output (Bishop, 1995).

Classification and regression are supervised learning tasks. Examples of unsupervised tasks are clustering, density estimation, and projection methods. According to Jain et al. (2000), data mining is one of many applications of pattern recognition. For other pattern recognition and machine learning tasks, see Bishop (2006).

3.2 Fields related to analysis of environmental data

3.2.1 Environmental statistics

Statistical data analysis is an important tool in environmental sciences. A traditional branch of statistics, biostatistics or biometry (also biometrics) is the application of statistics to a wide range of topics in biology.

Environmental statistics or environmetrics is the application of statistics and development of statistical methods for data from environmental studies. The journal Environmetrics (2007), e.g.,

> *publishes refereed papers on the development and application of quantitative methods in the environmental sciences.*

El-Shaarawi and Piegorsch (2002) define environmetrics simply as:

> *the use of measurements in the analysis, modeling, interpretation, and prediction of environmental phenomena.*

Environmental statistics uses both classical statistical methods and advanced modeling techniques in the analyses. Environmental statistics involves, e.g., the use of regression and time series models, risk assessment, and sampling strategies in environmental research (Piegorsch and Bailer, 2005). Details about sampling methods, models, and drawing conclusions in environmental sciences are presented by Manly (2001). For concrete examples of research problems in environmental statistics, see Piegorsch et al. (1998).

It is important to note that causality cannot be deduced from statistical dependencies in the data alone (Cherkassky and Mulier, 1998). This is especially important with ecological data because often there is no true replication and natural and anthropogenic disturbances are common as well as other confounding factors with poorly known consequences (Michener, 2000c).

3.2.2 Environmental informatics

In some cases, the traditional methods of environmental statistics do not provide the best results. It may be beneficial to use information technology and advanced data analysis methods for the difficult environmental problems. This branch of applied informatics is called environmental informatics (Avouris and Page, 1995) or enviromatics (also environmatics). Kolehmainen (2004) defined environmental informatics for the purposes of his thesis as follows:

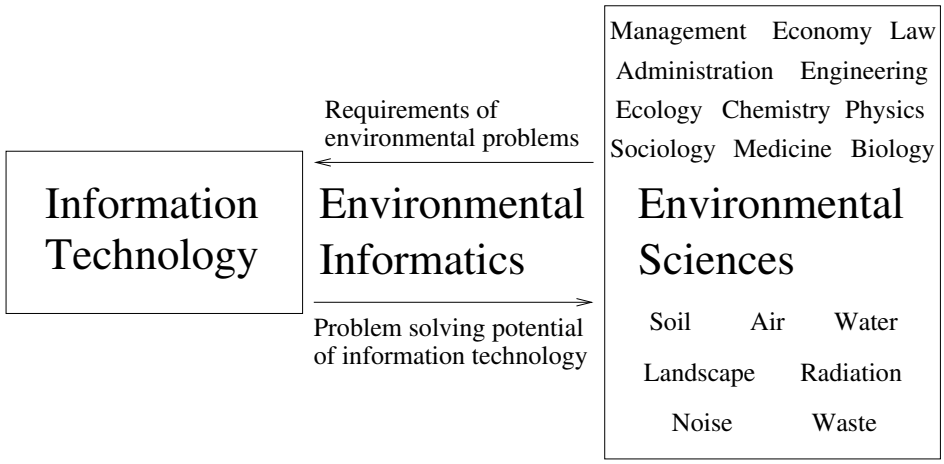


Figure 3.3: Environmental informatics resides between information technology and environmental sciences. The figure is modified from Hilty et al. (1995) and Page and Rautenstrauch (2001).

Environmental informatics is based on applying information technology to environmental issues using data-driven methods.

Green and Klomp (1999) see environmental informatics as a new paradigm, which states that local decisions and priorities should be set in a global context, and this can be achieved by setting up and using large-scale environmental information systems and databases. Le Duc (1996) notes that environmental informatics

can be decomposed into many areas – the common denominator being the combination of information technology and its use for environmental work.

The position of environmental informatics between information technology and environmental sciences is illustrated in Figure 3.3.

Typical problems for environmental informatics include many different non-homogeneous information sources (Thuvander, 2002), geographically coded data, and multidimensional data. In addition to analysis of measurement data, a significant concern in environmental informatics is handling of vague, uncertain, and incomplete knowledge (Avouris and Page, 1995; Cherkassky et al., 2006).

Ecological informatics or ecoinformatics is similar to environmental informatics, i.e., application and development of information technology and advanced data analysis methods for ecological studies (Recknagel, 2003). Ecology, however, is a sub-field of both environmental sciences and biology.

There are two particularly important paradigms in ecology: the balance of nature and the flux of nature. The first one is a classical paradigm and the second is a new one. The basic assumptions differ concerning, e.g., how common equilibrium points, disturbances, and human influences are in ecological systems. These differences affect the hypotheses and choice of models in ecological research (Michener, 2000b).

Another related field is bioinformatics, which is also a mediator between biology and information sciences. Despite the name, it is usually understood to concern mostly molecular biology. Therefore, environmental informatics and ecological informatics are not subfields of bioinformatics. However, the fields are strongly related with regard to the goals and the cooperation between biologists and information scientists or statisticians.

In environmental and ecological studies, it is often important that the constructed models can be inspected, modified, used, and verified by the domain experts. Therefore, according to Džeroski (2001), instead of using neural networks or other methods, which often produce black-box models, it may be beneficial to use symbolic machine learning methods, e.g., decision trees for classification and regression, algebraic, differential, and partial differential equations, etc. Džeroski (2001) also summarizes various applications of symbolic machine learning to ecological modeling.

Usually the domain experts are familiar with descriptive, behavioral, or physically-based process models of the environment. Therefore, introducing data-driven modeling may sometimes be a challenging task; e.g., the models may be criticized for not reflecting the physics of the modeled process (Cherkassky et al., 2006).

3.3 Models and methods in data analysis

3.3.1 Model complexity

When building a model, it is important to keep in mind that the model should give a good representation of the real system and not only the data used for training. That is, generalization should be the goal of modeling. The principles presented in this section describe different ways of approaching this goal.

Occam's razor—despite the name not invented by William of Ockham (Thorburn, 1918)—is a principle, which in case there are multiple models performing equally well, states that the simplest one should be chosen. If a model is too complex, it may produce very good results with the training data but not perform well with other data. This situation is called overfitting (Duda et al., 2001).

When considering the mean-square deviation between the true values and the model output, the deviation can be divided into two parts: bias and variance. Bias means how accurately on average the true function F can be estimated using

training data D. Variance means how much the estimate of F changes as the training set varies. The bias-variance dilemma (Geman et al., 1992; Heskes, 1998) tells that models with high flexibility, e.g., with a high number of parameters, tend to have low bias but high variance, whereas models with low flexibility often have higher bias and lower variance. In order to decrease both bias and variance, the size of the training set must be increased (Haykin, 1999).

One difficulty with multidimensional data is the curse of dimensionality, which means that the higher the dimension the more data is needed to find accurate parameter estimates. It has been demonstrated by Silverman (1986) that in case of density estimation, the number of observations needed for an accurate estimate grows extremely quickly as the dimension increases. Variable selection (Section 3.3.7) and projection methods can be used to fight this phenomenon.

According to Huber (1981):

> *Robustness signifies insensitivity to small deviations from the assumptions.*

A simple method to assess the generalization capability of a model and select its complexity is cross-validation (Stone, 1974). In m-fold cross-validation, the data set is divided into m disjoint sets of equal size. The model is trained m times and each time one of the sets is held out as a validation set. The generalization performance of the model is estimated as the mean of the m errors. In leave-one-out cross-validation, only one observation is used for validation at a time. Leave-one-out cross-validation is thus m-fold cross-validation, where m is equal to the number of observations.

In addition, different information criteria can be used to control the model complexity. Some information criteria, i.e., Mallows (1973) C_p criterion, minimum description length (MDL) information criterion, and Akaike's information criterion (AIC) (Hansen and Yu, 2001) are reviewed in Publication 5.

3.3.2 Clustering

Clustering is one common way of doing descriptive modeling. Other ways include density estimation and segmentation. As the name suggests, descriptive modeling aims to describe or summarize the data in a convenient form (Hand et al., 2001). Clustering can be characterized as unsupervised classification (Theodoridis and Koutroumbas, 1999). The goal is to divide the data set into natural classes or clusters without knowledge of any true classes. Typically, observations within a cluster should be similar to each other and observations in different clusters should be different from each other.

Clustering can be viewed as an ill-posed problem, i.e., any data set can be clustered in different ways with no clear criterion for preferring one clustering over another

(Angelini et al., 2007). The result of clustering depends, e.g., on the definition of similarity or distance between clusters and between observations. The choice of clustering algorithm should, therefore, match the requirements of the specific problem and the results should be verified using other methods. For more details and examples of difficult cases for clustering, see Pakkanen (2006).

One way of classifying clustering algorithms is to divide them into partitive and hierarchical algorithms (Theodoridis and Koutroumbas, 1999). Partitive algorithms divide the data into partitions, so that each observation belongs to only one of them. Hierarchical algorithms build a hierarchy of clusters (Everitt et al., 2001). The structure of a full cluster hierarchy is such that all the data belongs to the top level (root) cluster and at the bottom level each observation forms a separate cluster. In the intermediate levels, clusterings with different numbers of clusters may be chosen. Probably the best known partitive clustering algorithm is the k-means algorithm (MacQueen, 1967). Single linkage, complete linkage, and average linkage are traditional hierarchical clustering methods (Sneath and Sokal, 1973). Jain et al. (1999) have presented an overview and taxonomy of clustering methods. Also, Himberg (2003) provides an overview and many references.

The validity of the clustering result can be measured, e.g., by comparing the distances between observations in the same cluster to distances between clusters. A well-known cluster validity index of this type is the Davies-Bouldin index (Davies and Bouldin, 1979). For a survey on cluster validity indices, see Milligan and Cooper (1985) and for a more recent comparison, Bezdek and Pal (1998). In case the true class labels of the data are known, the quality of clustering can be measured using, e.g., mutual information (Bishop, 2006).

3.3.3 Self-organizing map

The self-organizing map (SOM) (Kohonen, 1981, 1982, 2001) is a useful tool in exploratory data analysis. It projects multidimensional data into a low-dimensional grid, which is easy to visualize. The SOM has been used successfully in numerous applications (Kaski et al., 1998; Oja et al., 2003).

The SOM consists of a regular, usually two-dimensional, grid of map units. Each unit i on the two-dimensional grid also has a d-dimensional prototype vector $\mathbf{m}_i$, where d is the dimension of the observations $\mathbf{x}_j$, $j = 1, \ldots, n$. Thus, the SOM defines a nonlinear projection from the d-dimensional data space to the two-dimensional grid. The SOM is trained to represent the original data by adapting its prototype vectors according to the distribution of the data set. The observations are mapped to map units with the closest prototype vector (the best-matching unit). Thus, in addition to nonlinear projection, the SOM also performs vector quantization. This representation can be used for visualization, clustering, and exploration of data (Kohonen, 2001).

Before training, the number of map units and the structure of the grid in the SOM are defined. The dimension of the prototype vectors is determined by the dimension of the data set. After initializing the map randomly or along the two greatest eigenvectors of the data, the training proceeds iteratively. At each training step t an observation $\mathbf{x}_j$ is first mapped to a map unit by looking for the best-matching unit c_j using a Euclidean distance measure between the observation and the set of map units.

$$c_j = \arg\min_i \|\mathbf{x}_j - \mathbf{m}_i(t)\| \tag{3.1}$$

Second, the prototype vectors are adapted to better represent the distribution of the data

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t) h_{c_j,i}(t) \left[\mathbf{x}_j - \mathbf{m}_i(t)\right], \tag{3.2}$$

where $\alpha(t)$ is a learning-rate factor and $h_{c_j,i}(t)$ is the neighborhood function. It is often a Gaussian function

$$h_{c_j,i}(t) = \exp\left[-\frac{\|r_{c_j} - r_i\|^2}{2\sigma^2(t)}\right], \tag{3.3}$$

where r_i is the location of the map unit i on the grid and $\sigma(t)$ corresponds to the width of the Gaussian function. Both $\alpha(t)$ and $\sigma(t)$ decrease during training. The original sequential training algorithm adapts the prototype vectors after each observation, whereas the batch training algorithm adapts the prototype vectors after all the data have been gone through (Kohonen, 2001).

The observations $\mathbf{x}_j$ often need to be normalized linearly before training, e.g., so that the mean of each variable is 0 and the variance is 1. The method used to normalize the data defines the distance between multidimensional vectors. For example, how should a change in N concentration measured as percent by mass be related to a change in mass per needle measured as grams? Normalizing all the variances to unity solves this problem by defining that changes in different variables are equal if they are in equal proportion to their standard deviations. As a result, all variables have equal weights in this sense.

The map units are connected to neighboring units on the grid by the neighborhood function. Therefore, the mapping from the original data space to the two-dimensional grid tends to preserve topological relationships. This means that observations close to each other in the data space tend to map to the same or close-by map units in the grid. How well the original topology is preserved by the SOM depends on the structure of the data (Martinetz and Schulten, 1994). Without the neighborhood function the SOM algorithm reduces to k-means clustering algorithm (Kohonen, 2001).

Conceptually, the SOM and its map units form an elastic net in the data space. This makes visualization of the grid useful in exploring the relationships of variables and the possible cluster structure of the data. The map can be visualized using component planes, each of which shows the values of one of the original d variables as colors on the grid. In addition, the map can be visualized with the

unified distance matrix (U-matrix) (Ultsch and Siemon, 1990), which shows the within-unit distances and distances between neighboring units on the grid.

The quality of the map can be measured with the quantization error, which is the average distance between each observation and its best-matching unit. The quantization error can be decreased by increasing the number of map units but this has the drawback that the number of observations per map unit decreases, which may lead to overfitting (Lampinen and Kostiainen, 1999). In addition to quantization, the topology preservation of the projection can be measured with the topographic error (Kiviluoto, 1996). It is defined as the percentage of observations for which the best-matching unit and the second-best-matching unit are not neighboring units on the grid. Also, the decomposition of the distortion measure can be used to assess the quantization quality and topological quality of the map (Vesanto et al., 2003).

3.3.4 Clustering of the self-organizing map

The principles of clustering of the SOM have been presented by Vesanto and Alhoniemi (2000). The U-matrix is a commonly used method for visual clustering of the data. A drawback is that the data analyst has to do the clustering manually based on the visualization, i.e., the result depends on the analyst. Therefore, an automated clustering approach is needed.

In Publication 1, an automated clustering algorithm based on the SOM is proposed. The author took part in designing the algorithm with Vesanto. The algorithm contains three phases:

1. The map units of the SOM are divided into so-called base-clusters using a region growing algorithm.
2. An agglomerative algorithm is used to build a hierarchy based on the base-clusters.
3. A pruning algorithm is used to remove extra intermediate clusters from the hierarchy.

The results of the proposed algorithm are found to be similar to the U-matrix. The algorithm is also compared with two other algorithms for clustering of the SOM: k-means clustering of the map units and a distance-matrix based method (Vellido et al., 1999). The experiments in the publication were designed together with Vesanto and the author conducted most of them. The proposed algorithm is found to produce on average better clustering results with an artificial data set than two other algorithms.

The clustering algorithm is used for analysis of forest nutrition data in Publications 2, 3, and 4. These analyses are described in Section 4.1.

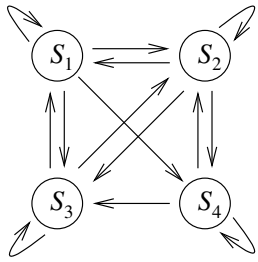


Figure 3.4: Possible transitions of a Markov model with four states visualized as a directed graph. The nodes represent the states and the arrows show all possible transitions in the model. Direct transitions from state 3 to state 4 and from state 4 to state 1 are not possible in this example. The other transitions have positive probabilities (not shown).

3.3.5 Markov chains

Markov chains are discrete-time stochastic processes, which can be used to describe the development of a system in time (Hamilton, 1994). A Markov chain is a series of random variables S_t. The values of S_t can assume only an integer value $1, \ldots, l$ and they can be interpreted as the states of a system at different time steps. A Markov chain has the Markov property, i.e., the conditional probability of the present state depends only on the previous state.

$$P(S_t = j | S_{t-1} = i, S_{t-2} = q, \ldots) = P(S_t = j | S_{t-1} = i) = p_{i,j} \tag{3.4}$$

The transition probability $p_{i,j}$ is the probability that state i will be followed by state j. The conditional probabilities of the present state given the previous state can be arranged in a $l \times l$ transition matrix, where l is the number of states.

$$\mathbf{A} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,l} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,l} \\ \vdots & \vdots & \ddots & \vdots \\ p_{l,1} & p_{l,2} & \cdots & p_{l,l} \end{bmatrix} \tag{3.5}$$

The conditional probabilities of a transition matrix can also be visualized as a directed graph. An example of a Markov model with four states visualized in this way is shown in Figure 3.4. In the example, $P(S_t = 4 | S_{t-1} = 3) = 0$ and $P(S_t = 1 | S_{t-1} = 4) = 0$ indicate impossible state transitions.

Hollmén et al. (1999) have presented a method for clustering Markov models or other probabilistic models using the SOM. If the states of the Markov model are not observed but have to be estimated using other observations, the model is called a hidden Markov model (Rabiner, 1989). Hidden Markov models have been used in a high number of machine learning applications (Bengio, 1999).

3.3.6 Linear regression

Regression is a predictive modeling task. The aim is to predict the value of a response variable using some regressor variables. In linear regression, the response is assumed to depend linearly on the regressors (Neter et al., 1996). The use of linear models can be justified by the fact that they are straightforward to interpret and that over short ranges, any process can be well approximated by a linear model (Guthrie et al., 2005). For nonlinear regression, e.g., generalized linear models (McCullagh and Nelder, 1989) and neural networks (Bishop, 2006) have been widely used.

The regression problem is to predict the values of a response variable Y using a number of d regressor variables X_i, $i = 1, \ldots, d$. The form of the available data is the following:

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \ldots & y_n \end{bmatrix}^T \tag{3.6}$$

and

$$\mathbf{x}_i = \begin{bmatrix} x_{i,1} & x_{i,2} & \ldots & x_{i,n} \end{bmatrix}^T. \tag{3.7}$$

Dependencies between the variables can be analyzed using the multiple linear regression model

$$y_j = \beta_0 + \beta_1 x_{1,j} + \ldots + \beta_d x_{d,j} + \epsilon_j, \tag{3.8}$$

where β_i, $i = 0, \ldots, d$ are the regression coefficients and error terms ϵ_j, $j = 1, \ldots, n$ are independent normally distributed random noise with zero mean and either known or unknown variance $N(0, \sigma_j^2)$. Equation 3.8 can equivalently be represented in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.9}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{d,1} \\ 1 & x_{1,2} & \ldots & x_{d,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \ldots & x_{d,n} \end{bmatrix}, \tag{3.10}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \ldots & \beta_d \end{bmatrix}^T, \tag{3.11}$$

and

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \ldots & \epsilon_n \end{bmatrix}^T. \tag{3.12}$$

In case the variance σ_j^2 of the Gaussian noise model can be assumed to be constant for $j = 1, \ldots, n$, the maximum likelihood estimates of the coefficients β_i are obtained by minimizing the residual sum of squares between the target values and the estimated values. The ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.13}$$

The coefficient of determination R^2 can be used to measure how well the regression model explains the observed data (Hair et al., 2006). It is the proportion of variability in the data set that is accounted for by the model

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}, \tag{3.14}$$

where

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1,j} + \ldots + \hat{\beta}_d x_{d,j} \tag{3.15}$$

is the prediction of the model and

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \tag{3.16}$$

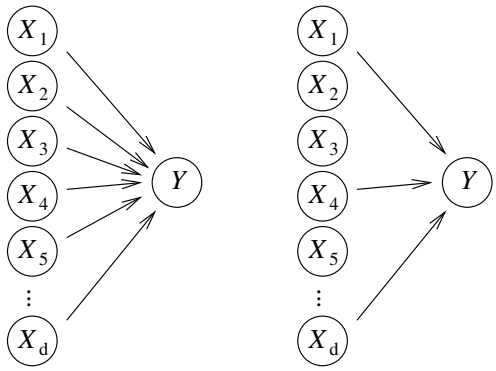
is the mean of the response.

The OLS regression model is, however, not always adequate for analysis of observed data. One difficulty, e.g., is that the number of regressors may be too high to estimate useful models (Publication 5). Another difficulty is that the quality of the observations is not necessarily constant, which is a basic assumption of the model (Publications 6 and 7). In addition to the publications, solutions for these problems are discussed in Sections 3.3.7 and 3.3.8.

3.3.7 Variable selection and sparse regression models

The full regression model estimated using OLS is not always the best model for prediction and interpretation of the dependencies. In case the number of regressors is high, the OLS models can be difficult to interpret. Also, if the number of observations is low, there is a risk of obtaining an overfitted model. In these kinds of cases, instead of using the full OLS model, it is often more sensible to use a sparse (or parsimonious) regression model.

In a linear sparse regression model, some of the regression coefficients β_i are set to zero. Selecting which variables are included in the model is called variable selection, input selection, or subset selection. According to Hair et al. (2006), the ratio of observations to regressors should preferably be at least between 15 to 20 to avoid overfitting. Setting coefficient values of non-informative regressors to zero increases the ratio of observations to regressors. In case of a high number of regressors, the most significant regressors are also more clearly seen from a sparse model. In addition, applying shrinkage to the regression coefficients can improve the prediction accuracy (Copas, 1983). A visual comparison of a full regression model and a sparse regression model is presented in Figure 3.5. The value of the response in the full model depends on the values of all regressors, whereas in the sparse model the value of the response depends only on the values of some regressors.



(a) Full model (b) Sparse model

Figure 3.5: A full (a) and a sparse (b) regression model. The arrows denote the dependencies between regressors X_i, $i = 1, \ldots, d$ and the response Y. In the sparse model, the value of the response does not depend on the values of all regressors.

Various methods for building sparse regression models by means of regression coefficient shrinkage and and subset selection are summarized in Publication 5. These include forward selection (Hastie et al., 2001), ridge regression (Hoerl and Kennard, 1970), least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), least angle regression (LARS) (Efron et al., 2004), the Curds & Whey (C&W) procedure (Breiman and Friedman, 1997), the nonnegative (nn) garrote (Breiman, 1995), and multivariate adaptive regression splines (MARS) (Friedman, 1991). In Publication 5, sparse regression models are also used for predicting the development of foliar nutrient concentrations in needles of foliar age class $C + 1$. This is explained in more detail in Section 4.2.

3.3.8 Trend detection and weighted regression models

Regression models can be used for detecting trends in time series. In this case, the regressor of interest is time. Using a linear OLS regression model for trend analysis is not justified if the precision of the observations is not constant because homoscedasticity, i.e., that the variance σ_j^2 of the noise is constant for $j = 1, \ldots, n$, is one of the basic assumptions of the model. A weighted regression model (Neter et al., 1996), however, is an effective trend analysis method for heteroscedastic data. The weight w_j of each observation is defined as the inverse of the noise variance. Thus, the method gives weights to observations according to their uncertainty

$$w_j = \frac{1}{\sigma_j^2}. \tag{3.17}$$

The weights can be arranged in a diagonal matrix as follows:

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \tag{3.18}$$

The weighted least squares (WLS) estimators of the regression coefficients are

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \tag{3.19}$$

where the response and regressor variables are as in Equations 3.6 and 3.10. The statistical significance of $\hat{\beta}_i \neq 0$ can be evaluated using the F-test (Jørgensen, 1993). A trend is detected if the null hypothesis that the regression coefficient corresponding to the time variable is zero, can be rejected with the chosen significance level.

Trend detection in environmental time series using weighted regression models is done in Publications 6 and 7. These analyses are also discussed in Section 4.3.

Chapter 4

Finding changes in mineral composition of forest foliage

This work can be characterized as environmental informatics as depicted in Figure 3.3. Tools of information technology are used and developed with the requirements of environmental problems. The work with regression models also fits under the scope of environmental statistics.

The whole process explained in this thesis can be considered knowledge discovery in environmental databases. Figure 3.1 shows the phases of the general KDD framework. The phases are also present in this study. First, a part of the forest nutrition, quality, and other data is selected for further investigation. Second, the target data is preprocessed and transformed: missing values are treated, scales of variables are normalized, noise information is collected, etc. Third, data mining is performed to the preprocessed and transformed data. The analyses of foliar nutrient concentration data using the SOM and regression models may be called data mining. The aim is to analyze the data to find unsuspected relationships and to summarize the data in understandable and useful ways (cf. Section 3.1.3). The amount of data is not exceptionally large but it is high dimensional and it is in a complex format due to the multiple sources. Finally, domain experts provide interpretation and evaluation of the results.

In the early phases of the work, the data were not well known. Therefore, a natural solution was to perform exploratory analysis of the data. As the research progressed, more data was collected when needed and hypotheses were generated and tested.

4.1 Exploratory analysis using the self-organizing map

The SOM has been used in various environmental data analysis problems. According to Park and Chon (2007), among unsupervised learning techniques, the SOM has been most widely used in extracting information from ecological data. The use of the SOM for data exploration in environmental informatics has been studied in the doctoral thesis of Kolehmainen (2004). His conclusion was that the SOM combined with Sammon's mapping (Sammon, 1969) has great potential in data exploration. Chon et al. (1996) compared the SOM with hierarchical clustering methods in analysis of macroinvertebrate communities in stream ecosystems and found that the low-dimensional presentation of the patterns makes the results comprehensible. In a similar application, Park et al. (2004) found hierarchical clustering of the SOM useful for assessing ecosystem quality and effects of environmental disturbances. Aguilera et al. (2001) have done water quality assessment based on nutrient data using the SOM and hierarchical clustering. In addition, clustering of the SOM has been used in many other environmental applications including nest-site selection of Black-tailed Gulls (Lee et al., 2006), analyzing riverbed habitats of mayflies (Hanquet et al., 2005), and environmental assessment of cumulative impact of multiple stressors on a large area (Tran et al., 2003). For more applications of the SOM and other machine learning methods in ecological informatics, see Park and Chon (2007).

4.1.1 Clustering of forest nutrition data

Exploratory analysis of forest nutrition data using the SOM is performed. First, four variables describing the foliage, which are considered ecologically and physiologically most important by forest scientists Luyssaert and Raitio, are selected for further analysis. Next, a SOM is trained with these variables. The U-matrix of the SOM suggests that there is cluster structure in the data (Publication 2). Then, clustering of the SOM is performed with the algorithm proposed in Publication 1 (see also Section 3.3.4). A cluster hierarchy is obtained and a clustering with six clusters is selected from the hierarchy for further analysis. The progress of the analysis is shown in Figure 4.1. The results of these analyses are presented in Publications 2, 3, and 4. Some results of the analysis of the forest nutrition data using spatial statistics (Ripley, 1981), clustering of the SOM, and hidden Markov models have been presented in the master's thesis of the author (Sulkava, 2003).

Values of some missing observations in data used in Publication 2 were determined by Metla for analyses of Publications 3 and 4. This causes the slight difference in the clustering results. Stands with similar nutrition profiles form their own cluster. The six clusters are characterized by the average values of all the measured variables within the clusters. These group nutrition profiles are used to

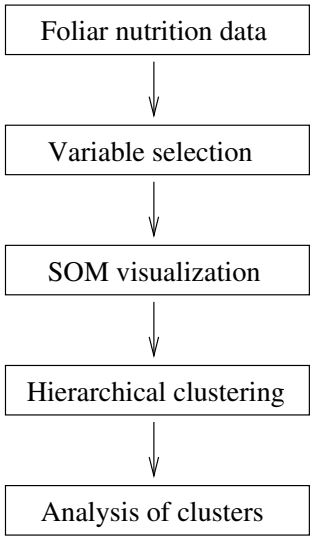


Figure 4.1: Analysis of forest nutrition data using clustering of the SOM.

characterize changes in the mineral composition of the foliage and assess how it is affected by N and S deposition.

4.1.2 Temporal models of forest nutrition data

The SOM and its clustering, as such, give a static view on the data. The next step is to study if there are some kind of temporal changes in the data. This is done based on the clustering. The clusters are used as states of the foliage.

The time dimension is not directly used in the clustering. Instead, the cluster sequence is used as a Markov chain, and the temporal behavior of the foliage is studied with cluster transition matrices (Publications 2, 3, and 4) and graphs (Publication 2) as presented in Section 3.3.5. When the clustering is viewed this way, the solution is also related to so-called (k, h)-segmentation (Gionis and Mannila, 2003), where k is the number of segments and h is the number of states. The difference is that in the clustering approach, k is not limited. For more information on time series segmentation, see Terzi (2006).

Our experiences with using the SOM for exploratory analysis of environmental data are positive. The visualization of the map with the component planes and U-matrix were found to be understandable to the domain experts. Also, the results of clustering and cluster transitions are effective tools for hypothesis generation.

4.2 Sparse regression models for development of forest nutrition

The number of possible regressors is often high in environmental sciences. Sometimes, the importance of the regressors can be estimated based on some a priori knowledge but in many cases this is not possible. Therefore, the use of variable selection and sparse regression methods is important.

Svenning and Skov (2005), e.g., have used sparse regression methodology to analyze how much tree species composition is related to climate, other environmental factors, and history. In addition, Reineking and Schröder (2006) have studied the use of ridge, lasso, and other methods for regularization and subset selection in a simulation study of habitat regression models. The selected subsets were not always the correct ones, but regularization was found to improve the predictive performance. Also, Ramadan et al. (2001) found that stepwise variable selection methods improved the performance of soil source prediction using microbial community DNA (deoxyribonucleic acid) data.

There are many needle year-classes in coniferous trees. Because there were measurements of both the current C and previous $C+1$ year needles in the data studied in this thesis, it is possible to study what kinds of factors contribute to the aging of the needles between two consecutive years, i.e., how the mineral composition of the needles changes as they turn from foliar age class C to $C+1$ during one year.

In Publication 5, the objective is to predict and explain the nutrient concentrations and mass of needles in foliar age class $C+1$ using weather and other environmental measurements and the values of foliar age class C one year earlier, i.e., in practice the same needles in the previous year. This problem has also been studied earlier by Sulkava et al. (2004). The aim is to build models, which in addition to good prediction capabilities, are also easy to understand. Linear sparse regression models are constructed separately for each response. The modeling task is visualized in Figure 4.2. If the only goal was the prediction accuracy, it might have been beneficial to use, e.g., some nonlinear or piecewise linear procedure.

In Publication 5, four kinds of regression models are compared: full multiple regression models, simple regression models, and sparse regression models estimated using two algorithms, i.e., LARS and forward selection. The quality of the models is validated using cross-validation. The sparse models are found to meet the aims of the study well. The prediction accuracy of the sparse models is similar to the full models and often clearly better than the prediction accuracy of the simple regression models. The sparse models are, however, much less complex than the full models. Therefore, they are more useful in assessing the importance of the numerous regressors.

The result is that different factors are important in explaining the concentrations of different nutrients in the $C+1$ needles. It is noted, that some of the nutrient

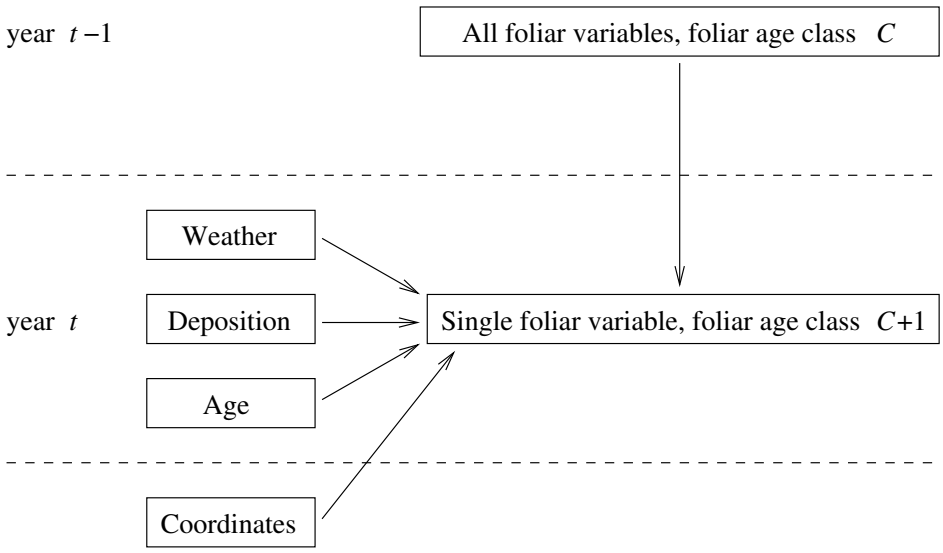


Figure 4.2: Modeling changes in needles between two years. All foliar variables of foliar age class C in year $t-1$ and environmental factors in year t are used to predict a foliar variable of foliar age class $C+1$ in year t.

concentrations in $C+1$ needles contain more information than others, e.g., Ca concentration in the needles does not change much between two consecutive years, whereas N concentrations in C and $C+1$ needles of two years are less connected. This information may be helpful in deciding which measurements are important in the monitoring program (Luyssaert, S., personal communication, April, 2005).

4.3 Laboratory quality and trend detection

Trend is an important concept in environmental sciences. It is a simple and understandable way of characterizing changes in the monitored part of the environment. Due to constant changes in environment, trends can be found in many kinds of environmental time series.

The power of a test is the probability that a false null hypothesis is rejected. Power analysis (Cohen, 1988) is important for designing experiments and assessing how likely monitoring programs will detect changes of a certain magnitude in environment.

Gerrodette (1987) has done power analysis of unweighted regression models for trend detection in heteroscedastic environmental time series data. He quantified how the power of a trend test is affected by the precision and number of observa-

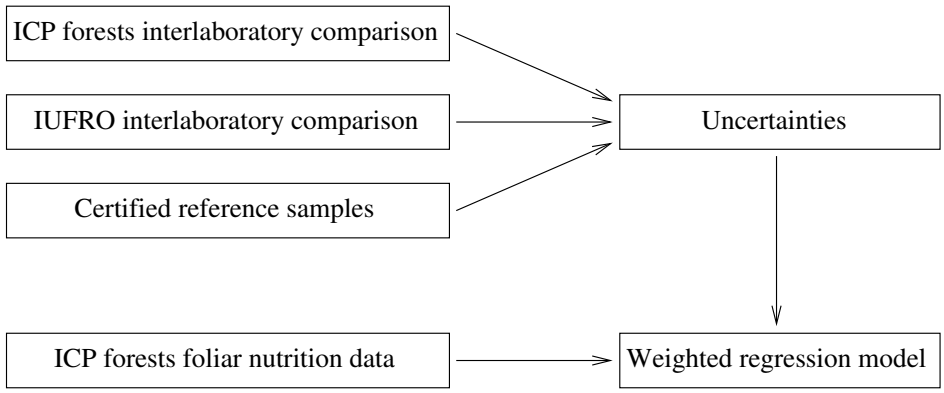


Figure 4.3: Fusion of measurement quality metadata from three different sources and forest nutrition data makes it possible to use weighted regression models for trend detection.

tions. Fryer and Nicholson (1999) compared locally weighted regression smoothers to other trend tests and found that a parametric test, e.g., linear regression for a linear trend, is more powerful than the smoothers and nonparametric tests. Hess et al. (2001) have done a comparison of seven methods for trend detection in environmental data. They found that a seasonal Kendall test and a t-test adjusted for seasonality have the highest power among the tests. Weatherhead et al. (1998) have calculated how the magnitude of noise and its autocorrelation affect the time needed to detect a linear trend in environmental data. Detecting trends in environmental data using regression models has also been studied in the doctoral thesis of Hussian (2005). In addition, Erästö (2006) has used smoothing methods for trend detection in exploratory data analysis of environmental time series data. Clarke (2002a) has studied how an iterative reweighted least squares algorithm can be used to fit generalized linear models for trend detection in Gumbel-distributed hydrologic time series data. He found that the power of the procedure is higher than the power of linear regression and the Mann-Kendall test (Clarke, 2002b).

The effect of data quality on trend detection is studied in Publications 6 and 7. The difference from the trend detection studies discussed above is that weighted regression models are used for heteroscedastic data. Simulation is used to compute the power of the tests in different scenarios of changing data quality. It is computed, how the power of trend tests depends on the precision of the data and the length of time series. Different scenarios of changing precision are considered in both single measurements (Publications 6 and 7) and ratio of two measurements (Publication 7).

Additionally, laboratory quality test data is combined from three different sources containing information about repeated measurements of the same sample and thus, the uncertainty of the measurements (Figure 4.3). With these metadata, it

is possible to estimate the accuracy and precision of the laboratories analyzing the foliar samples. Fusion of various quality data and foliar nutrient concentration data makes it possible to use weighted regression models to determine whether there are statistically significant trends in N and S concentrations and Ca/Al concentration ratio. The observed data is modeled with a simple weighted linear regression model in Publication 6. A more realistic iterative reweighted model including more sources of variation is used in Publication 7. Also, more data was obtained for experiments of Publication 7.

Chapter 5

Summary and conclusions

As a conclusion, it can be said that modern data analysis methods are useful tools in environmental informatics and environmental statistics. Good methods are understandable for the environmental scientists and at the same time reliable, robust, and helpful for discovering important relationships in the data.

In cooperation between environmental scientists and information scientists, what makes the relationship flourish is the knowledge of both sides about their field and efficient communication concerning the specific needs of a certain problem and the properties of the methods. Without these ingredients, the results of cooperation projects may not be satisfactory.

In this thesis, a number of data analysis methods were developed and applied to real problems in environmental sciences or more specifically forest science. These methods can also be used in similar problems of other branches of environmental sciences. Depending on the problem, the methods may need to be tuned to meet the specific requirements of the problem.

First, an algorithm for automatic clustering of the self-organizing map was proposed. It was found to produce good results with an artificial data set. The use of the SOM for visualization and clustering were found to be advantageous methods in analysis of forest nutrition data. Also, using the clusters as states of a Markov chain made it possible to study the temporal changes in the data. These methods helped to gain insight about the condition and development of forests in Finland based on the initially not so well known data.

Second, sparse regression methodology was studied in the task of explaining which factors affect the aging of tree needles and how much. Four kinds of linear regression models were compared: full multiple regression models, simple regression models, and sparse regression models estimated using two algorithms. The sparse regression techniques were found to produce models, which are easy to under-

stand but whose prediction capabilities are still comparable to the full multiple regression models.

Third, power analysis of weighted regression models was conducted in order to quantify how much the quality of analytical methods in a laboratory affects the ability to detect statistically significant trends in heteroscedastic time series data. Laboratory quality data was collected from different sources and combined to estimate the precision of the laboratories of Metla. Noise models were built for different scenarios of laboratory quality and other variation of data in both single and combined indices. Based on the noise models, the effect of data quality on trend detection was quantified using simulation and observed forest nutrition data.

When selecting the tools for an environmental data analysis problem, it is often important that the domain experts understand what the methods produce as outputs. Very complex methods are not necessarily needed in all problems (Laine, 2003), but using very simple methods may not be the way to go either. Firstly, the environmental scientists have probably already tried them and either found them useful, in which case the problem has been solved, or found them not so useful, in which case there is no need to try again. Secondly, environmental scientists often have some auxiliary information—sometimes in the form of metadata—concerning the specific environmental problem, which should be taken into account in the analysis. Trying to find a suitable complexity of methods is not easy, but it has to be done in order to do successful research in environmental informatics and statistics.

Many challenges were met during this thesis work but they all could be dealt with. As a result, new knowledge was obtained concerning the functioning and analysis of forest ecosystems and the use of the studied data analysis methods in forest science. The importance of this work is that the data analysis methods were found useful in gaining new knowledge of forest ecosystems. The methods are understandable to the forest scientists and, therefore, make it possible to draw meaningful conclusions about changes in the environment and their possible causes.

The concept of nutrition profile was proposed for characterizing the nutritional status of foliage. The nutrition profile has the advantage over more traditional ways for characterizing the foliage especially in monitoring programs that it accounts for all element concentrations, contents, and interactions between two or more elements. Based on the analysis of the nutrition profiles, it was found that evidence for N and S deposition-induced changes in the mineral composition of tree foliage had decreased from late eighties to late nineties.

The sparse regression models for predicting the nutritional status of foliar age class $C + 1$ showed the importance of environmental factors for the nutritional status. The models also revealed interconnections between element concentrations, which can be used for guiding the laboratory work after the sample collection.

The analysis using weighted regression models for trend detection quantified the effect of data quality on detecting changes in environmental time series. The results show that if the variability of representative sampling is low, poor quality of the instrumental analysis can cause from years to decades-long delay in detecting changes in environmental monitoring or long-term ecological research programs.

This thesis does not contain all useful ways of studying the functioning of the environment. In the future, the work can be extended, e.g., by doing similar analyses to larger data sets of forest foliage, i.e., more variables, more stands, and more years. In addition, use of nonlinear models for the regression problems can be examined. Comparison of different trend tests in case of heteroscedastic time series is also one direction of future research.

Naturally, the research can also be continued with other environmental problems. This has already been done for a while in the analysis of net carbon-exchange of forests (Luyssaert et al., 2007) and automatic detection of onset and cessation of growing season of trees based on automated dendrometer data (Sulkava et al., 2007). Understanding the functioning of the environment better is important, and suitable tools of data analysis are helpful in this task. Therefore, their use in environmental science should be encouraged.

References

Aguilera, P. A., Garrido Frenich, A., Torres, J. A., Castro, H., Martinez Vidal, J. L., and Canton, M. (2001). Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. *Water Research*, 35(17):4053–4062.

Angelini, L., Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2007). Natural clustering: the modularity approach. *Journal of Statistical Mechanics: Theory and Experiment*.

Argyris, C. (1982). *Reasoning, Learning, and Action: Individual and Organizational*. Social and Behavioral Science Series and Management Series. Jossey-Bass, San Francisco, CA, USA, first edition.

Avouris, N. M. and Page, B., editors (1995). *Environmental Informatics: Methodology and Applications of Environmental Information Processing*, volume 6 of *Computer and information science*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Bengio, Y. (1999). Markovian models for sequential data. *Neural Computing Surveys*, 2:129–162.

Berthouex, P. M. and Brown, L. C. (2002). *Statistics for Environmental Engineers*. Lewis Publishers, Boca Raton, FL, USA, second edition.

Bezdek, J. C. and Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 28(3):301–315.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York, NY, USA.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, USA.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multivariate regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(1):3–54.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0 step-by-step data mining guide. Technical report, CRISP-DM consortium. `http://www.crisp-dm.org/CRISPWP-0800.pdf`.

Cherkassky, V., Krasnopolsky, 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.

Cherkassky, V. and Mulier, F. (1998). *Learning from data: concepts, theory and methods*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, New York, NY, USA.

Chon, T.-S., Park, Y. S., Moon, K. H., and Cha, E. Y. (1996). Patternizing communities by using an artificial neural network. *Ecological Modelling*, 90(1):69–78.

Clarke, R. T. (2002a). Estimating time trends in Gumbel-distributed data by means of generalized linear models. *Water Resources Research*, 38(7):16.1–16.11.

Clarke, R. T. (2002b). Fitting and testing the significance of linear trends in gumbel-distributed data. *Hydrology and Earth Systems Sciences*, 6(1):17–24.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, Hillsdale, NJ, USA, 2nd edition.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354.

Cumming, J. A. and Wooff, D. A. (2007). Dimension reduction via principal variables. *Computational Statistics & Data Analysis*, 52(1):550–565.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Dowdy, S., Wearden, S., and Chilko, D. (2004). *Statistics for Research*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, NJ, USA, third edition.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, New York, NY, USA, second edition.

Džeroski, S. (2001). Applications of symbolic machine learning to ecological modelling. *Ecological Modelling*, 146:263–273.

EANET (2007). Acid deposition monitoring network in East Asia. `http://www.eanet.cc/`.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

El-Shaarawi, A. H. and Piegorsch, W. W., editors (2002). *Encyclopedia of Environmetrics*. Wiley, Chichester, UK.

Emanuel, K. (2005). Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 436(7051):686–688.

EMEP (2007). Co-operative programme for monitoring and evaluation of the long-range transmission of air pollutants in europe. `http://www.emep.int/`.

Environmetrics (2007). Environmetrics. `http://www3.interscience.wiley.com/cgi-bin/jhome/6285`.

Erästö, P. (2006). *Studies in Trend Detection of Scatter Plots with Visualization*. PhD thesis, University of Helsinki, Helsinki, Finland.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster Analysis*. Arnold, London, UK, fourth edition.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

FFCMP (2007). The Finnish forest condition monitoring programme (Forest focus / ICP forests). `http://www.metla.fi/metinfo/metsienterveys/metsientila/ICP-forests/`.

FHM (2007). USDA forest service – national forest health monitoring program. `http://fhm.fs.fed.us/`.

FMI (2007). Finnish meteorological institute. `http://www.fmi.fi/en/index.html`.

Forest Focus (2007). Forest focus. `http://europa.eu/scadplus/leg/en/lvb/l28125.htm`.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1991). Knowledge discovery in databases: An overview. In Piatetsky-Shapiro, G. and Frawley, W. J., editors, *Knowledge Discovery in Databases*, chapter 1, pages 1–27. AAAI Press / The MIT Press, Menlo Park, CA, USA.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Fryer, R. J. and Nicholson, M. D. (1999). Using smoothers for comprehensive assessment of contaminant time series in marine data. *ICES Journal of Marine Science*, 56(5):779–790.

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

Gerrodette, T. (1987). A power analysis for detecting trends. *Ecology*, 68(5):1364–1372.

Gionis, A. and Mannila, H. (2003). Finding recurrent sources in sequences. In Vingron, M., Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the Seventh International Conference on Research in Computational Molecular Biology – RECOMB 2003*, pages 123–130., Berlin, Germany. Max Planck Institute for Molecular Genetics and Berlin Center for Genome Based Bioinformatics, ACM Press.

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1(1):11–28.

Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer, New York, NY, USA, third edition.

Green, D. G. and Klomp, N. I. (1999). Environmental informatics – a new paradigm for coping with complexity in nature. *Complexity International*, 6.

Günther, O. (1998). *Environmental Information Systems*. Springer, Berlin, Germany.

Guthrie, W., Filliben, J., and Heckert, A. (2005). *NIST/SEMATECH e-Handbook of Statistical Methods*, chapter 4. Process Modeling. National Institute of Standards and Technology. `http://www.itl.nist.gov/div898/handbook/`.

Hair, Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2006). *Multivariate Data Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, USA, sixth edition.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ, USA.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, USA.

Hanquet, D., Legalle, M., Compin, A., and Céréghino, R. (2005). Assessment of an artificial intelligence technique in investigating habitat partitioning by coexisting benthic invertebrates in gravel-bed rivers. *River Research and Applications*, 21(6):629–639.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Upper Saddle River, NJ, USA, second edition.

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433.

Hess, A., Iyer, H., and Malm, W. (2001). Linear trend analysis: a comparison of methods. *Atmospheric Environment*, 35(30):5211–5222.

Hilty, L. M., Page, B., Radermacher, F. J., and Riekert, W.-F. (1995). Environmental informatics as a new discipline of applied computer science. In Avouris, N. M. and Page, B., editors, *Environmental Informatics: Methodology and Applications of Environmental Information Processing*, volume 6 of *Computer and information science*, pages 1–11. Kluwer Academic Publishers, Dordrecht, Netherlands.

Himberg, J. (2003). *From insights to innovations: data mining, visualization, and user interfaces*. D.Sc. thesis, Helsinki University of Technology, Espoo, Finland.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (2000). *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York, NY, USA, Wiley classics library edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hollmén, J., Tresp, V., and Simula, O. (1999). A self-organizing map for clustering probabilistic models. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN99)*, number 470 in Conference Publication series, pages 946–951, Edinburgh, UK. The Informatics Division of the Institution of Electrical Engineers, IEE.

Huber, P. J. (1981). *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, NY, USA.

Hussian, M. A. E. H. (2005). *Monotonic and Semiparametric Regression for the Detection of Trends in Environmental Quality Data*. Doctoral thesis, Linköpings Universitet, Linköping, Sweden.

ICP Forests (2007). International co-operative programme on assessment and monitoring of air pollution effects on forests. `http://www.icp-forests.org`.

IUFRO (2007). International union of forest research organizations. `http://www.iufro.org/`.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

Jørgensen, B. (1993). *The theory of linear models*. Chapman & Hall, New York, NY, USA.

Kaski, S., Kangas, J., and Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1:102–350.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. In *Proceedings of the International Conference on Neural Networks (ICNN'96)*, volume 1, pages 294–299, Piscataway, NJ, USA. IEEE Neural Networks Council.

Kohonen, T. (1981). Automatic formation of topological maps of patterns in a self-organizing system. In Oja, E. and Simula, O., editors, *Proceedings of The Second Scandinavian Conference on Image Analysis*, pages 214–220, Helsinki, Finland. Pattern Recognition Society of Finland.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kohonen, T. (2001). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Germany, third edition.

Kolehmainen, M. T. (2004). *Data exploration with self-organizing maps in environmental informatics and bioinformatics*. D.Sc. thesis, Helsinki University of Technology, Espoo, Finland.

Laine, S. (2003). *Using visualization, variable selection and feature extraction to learn from industrial data*. D.Sc. thesis, Helsinki University of Technology, Espoo, Finland.

Lampinen, J. and Kostiainen, T. (1999). Overtraining and model selection with the self-organizing map. In *Proceedings of the International Joint Conference on Neural Networks – IJCNN'99*, volume 3, pages 1911–1915, Washington, DC, USA. IEEE.

Le Duc, M. (1996). *Constructivist Systemics. Theoretical Elements and Applications in Environmental Informatics*. Doctoral dissertation, Stockholm University, Stockholm, Sweden.

Lee, W.-S., Kwon, Y.-S., Yoo, J.-C., Song, M.-Y., and Chon, T.-S. (2006). Multivariate analysis and self-organizing mapping applied to analysis of nest-site selection in Black-tailed Gulls. *Ecological Modelling*, 193(3–4):602–614.

Liff, C. I., Riitters, K. H., and Hermann, K. A. (1994). Forest health monitoring case study. In Michener, W. K., Brunt, J. W., and Stafford, S. G., editors, *Environmental Information Management and Analysis: Ecosystem to Global Scales*, pages 101–112. Taylor & Francis, London, UK.

Luyssaert, S., Janssens, I. A., Sulkava, M., Papale, D., Dolman, A. J., Reichstein, M., Hollmén, J., Martin, J. G., Suni, T., Vesala, T., Lousteau, D., Law, B. E., and Moors, E. J. (2007). Photosynthesis drives anomalies in net carbon-exchange of pine forests at different latitudes. *Global Change Biology*, 13(10):2110–2127.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA, USA. University of California, National Science Foundation, National Institutes of Health, Air Force Office of Scientific Research, Army Research Office, Office of Naval Research, University of California Press.

Magnani, F., Mencuccini, M., Borghetti, M., Berbigier, P., Berninger, F., Delzon, S., Grelle, A., Hari, P., Jarvis, P. G., Kolari, P., Kowalski, A. S., Lankreijer, H., Law, B. E., Lindroth, A., Loustau, D., Manca, G., Moncrieff, J. B., Rayment, M., Tedeschi, V., Valentini, R., and Grace, J. (2007). The human footprint in the carbon cycle of temperate and boreal forests. *Nature*, 447(7146):848–852.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

Manly, B. F. J. (2001). *Statistics for Environmental Science and Management*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Martinetz, T. and Schulten, K. (1994). Topology representing networks. *Neural Networks*, 7(3):507–522.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, UK, 2nd edition.

Metla (2007). Finnish forest research institute. `http://www.metla.fi/index-en.html`.

Michener, W. K. (2000a). Metadata. In Michener, W. K. and Brunt, J. W., editors, *Ecological Data: Design, Management and Processing*, Methods in Ecology, chapter 5, pages 92–116. Blackwell Science, Oxford, UK.

Michener, W. K. (2000b). Research design: Translating ideas to data. In Michener, W. K. and Brunt, J. W., editors, *Ecological Data: Design, Management and Processing*, Methods in Ecology, chapter 1, pages 1–24. Blackwell Science, Oxford, UK.

Michener, W. K. (2000c). Transforming data into information and knowledge. In Michener, W. K. and Brunt, J. W., editors, *Ecological Data: Design, Management and Processing*, Methods in Ecology, chapter 7, pages 142–161. Blackwell Science, Oxford, UK.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, Chicago, IL, USA, 4th edition.

Oja, M., Kaski, S., and Kohonen, T. (2003). Bibliography of self-organizing map (SOM) papers: 1998–2001 addendum. *Neural Computing Surveys*, 3:1–156.

Page, B. (1995). Database technologies for environmental data management. In Avouris, N. M. and Page, B., editors, *Environmental Informatics: Methodology and Applications of Environmental Information Processing*, volume 6 of *Computer and information science*, pages 39–51. Kluwer Academic Publishers, Dordrecht, Netherlands.

Page, B. and Rautenstrauch, C. (2001). Environmental informatics—methods, tools and applications in environmental information processing. In Rautenstrauch, C. and Patig, S., editors, *Environmental Information Systems in Industry and Public Administration*. Idea Group Publishing, Hershey, PA, USA.

Pakkanen, J. (2006). *Approaches for Content-Based Retrieval of Surface Defect Images*. D.Sc. thesis, Helsinki University of Technology, Espoo, Finland.

Park, Y.-S. and Chon, T.-S. (2007). Biologically-inspired machine learning implemented to ecological informatics. *Ecological Modelling*, 203(1–2):1–7.

Park, Y.-S., Chon, T.-S., Kwak, I.-S., and Lek, S. (2004). Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. *Science of the Total Environment*, 327:105–122.

Phillips, D. C. and Soltis, J. F. (2004). *Perspectives on Learning*. Thinking About Education Series. Teachers College Press, New York, NY, USA, fourth edition.

Piegorsch, W. W. and Bailer, A. J. (2005). *Analyzing Environmental Data*. John Wiley & Sons.

Piegorsch, W. W., Smith, E. B., Edwards, D., and Smith, R. L. (1998). Statistical advances in environmental science. *Statistical Science*, 13(2):186–208.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ramadan, Z., Song, X.-H., Hopke, P. K., Johnson, M. J., and Scow, K. M. (2001). Variable selection in classification of environmental soil samples for partial least square and neural network models. *Analytica Chimica Acta*, 446:233–244.

Recknagel, F., editor (2003). *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation*. Springer, Berlin, Germany.

Reineking, B. and Schröder, B. (2006). Constrain to perform: Regularization of habitat models. *Ecological Modelling*, 193(3–4):675–690.

Ripley, B. D. (1981). *Spatial Statistics*. Wiley Series in Probability And Mathematical Statistics. John Wiley & Sons, New York, NY, USA.

Sammon, Jr., J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

Schulze, E.-D., Wirth, C., and Heimann, M. (2000). Climate change: Managing forests after Kyoto. *Science*, 289(5487):2058–2059.

Senge, P. M., Kleiner, A., Roberts, C., Ross, R. B., and Smith, B. J. (1996). *The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization*. Nicholas Brealey Publishing, London, UK.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, UK.

Similä, T. (2007). *Advances in variable selection and visualization methods for analysis of multivariate data*. D.Sc. thesis, Helsinki University of Technology, Espoo, Finland.

Slagle, R. L. (1994). Standards for integration of multisource and cross-media environmental data. In Michener, W. K., Brunt, J. W., and Stafford, S. G., editors, *Environmental Information Management and Analysis: Ecosystem to Global Scales*, pages 221–233. Taylor & Francis, London, UK.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. A series of books in biology. W. H. Freeman and Company, San Francisco, CA, USA.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, 36(2):111–147.

Sulkava, M. (2003). Identifying spatial and temporal profiles from forest nutrition data. Master's thesis, Helsinki University of Technology, Espoo, Finland.

Sulkava, M., Mäkinen, H., Nöjd, P., and Hollmén, J. (2007). CUSUM charts for detecting onset and cessation of xylem formation based on automated dendrometer data. In Horová, I. and Hřebíček, J., editors, *TIES 2007 – 18th annual meeting of the International Environmetrics Society, Book of Abstracts*, page 111, Mikulov, Czech Republic. The International Environmetrics Society, Masaryk University.

Sulkava, M., Tikka, J., and Hollmén, J. (2004). Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. In Džeroski, S., Ženko, B., and Debeljak, M., editors, *Proceedings of the Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, pages 57–58, Bled, Slovenia.

Svenning, J.-C. and Skov, F. (2005). The relative roles of environment and history as controls of tree species composition and richness in Europe. *Journal of Biogeography*, 32(6):1019–1033.

Terzi, E. (2006). *Problems and Algorithms for Sequence Segmentations*. PhD thesis, University of Helsinki, Helsinki, Finland.

Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press.

Thorburn, W. M. (1918). The myth of Occam's razor. *Mind*, 27(3):345–353.

Thuvander, L. (2002). *Towards Environmental Informatics for Building Stocks: A conceptual model for an Environmental Building Stock Information System for Sustainable Development – EBSIS^{SD}*. PhD thesis, Chalmers University of Technology, Göteborg, Sweden.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tran, L. T., Knight, C. G., O'Neill, R. V., Smith, E. R., and O'Connell, M. (2003). Self-organizing maps for integrated environmental assessment of the Mid-Atlantic region. *Environmental Management*, 31(6):822–835.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley, Reading, MA, USA.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25.

Ultsch, A. and Siemon, H. P. (1990). Kohonen's self organizing feature maps for exploratory data analysis. In *Proceedings of International Neural Network Conference (INNC'90)*, pages 305–308, Dordrecht, Netherlands. Kluwer.

Vellido, A., Lisboa, P. J. G., and Meehan, K. (1999). Segmentation of the online shopping market using neural networks. *Expert Systems with Applications*, 17(4):303–314.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600.

Vesanto, J., Sulkava, M., and Hollmén, J. (2003). On the decomposition of the self-organizing map distortion measure. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 11–16, Kitakyushu, Japan.

Weatherhead, E. C., Reinsel, G. C., Tiao, G. C., Meng, X.-L., Choi, D., Cheang, W.-K., Keller, T., DeLuisi, J., Wuebbles, D. J., Kerr, J. B., Miller, A. J., Oltmans, S. J., and Frederick, J. E. (1998). Factors affecting the detection of trends: Statistical considerations and applications to environmental data. *Journal of Geophysical Research*, 103(D14):17149–17161.