Linköping Studies in Statistics • No. 7 Doctoral Thesis
Linköping Studies in Arts and Science • No. 343 Doctoral Thesis

Monotonic and Semiparametric Regression for the Detection of Trends in Environmental Quality Data

Mohamed A. E. H. Hussian

LINKÖPINGS UNIVERSITET
FACULTY OF ARTS AND SCIENCES

Department of Mathematics, Division of Statistics
Linköpings Universitet
SE-581 85 Linköping, Sweden
Linköping 2005

Linköping Studies in Statistics • No. 7 Doctoral Thesis
Linköping Studies in Arts and Science • No. 343 Doctoral Thesis

At the Faculty of Arts and Science at Linköpings universitet, research and doctoral studies are carried out within broad problem areas. Research is organized in interdisciplinary research environments and doctoral studies mainly in graduate schools. Jointly, they publish the series Linköping Studies in Arts and Sciences. This thesis comes from the Division of Statistics at the Department of Mathematics.

Distributed by:
Department of Mathematics, Division of Statistics
Linköpings universitet
581 83 Linköping

Author: Mohamed Hussian

Title: Monotonic and Semiparametric Regression for the Detection of Trends in Environmental Quality Data

Edition 1:1
ISBN 91-85457-70-1
ISSN 0282-9800
Department of Mathematics, ISSN 1651-1700

Printed by Liu-Tryck, Linköping 2005

Abstract

Natural fluctuations in the state of the environment can long conceal or distort important trends in the human impact on our ecosystems. Accordingly, there is increasing interest in statistical normalisation techniques that can clarify the anthropogenic effects by removing meteorologically driven fluctuations and other natural variation in time series of environmental quality data. This thesis shows that semi- and nonparametric regression methods can provide effective tools for applying such normalisation to collected data. In particular, it is demonstrated how monotonic regression can be utilised in this context. A new numerical algorithm for this type of regression can accommodate two or more discrete or continuous explanatory variables, which enables simultaneous estimation of a monotonic temporal trend and correction for one or more covariates that have a monotonic relationship with the response variable under consideration. To illustrate the method, a case study of mercury levels in fish is presented, using body length and weight as covariates. Semiparametric regression techniques enable trend analyses in which a nonparametric representation of temporal trends is combined with parametrically modelled corrections for covariates. Here, it is described how such models can be employed to extract trends from data collected over several seasons, and this procedure is exemplified by discussing how temporal trends in the load of nutrients carried by the Elbe River can be detected while adjusting for water discharge and other factors. In addition, it is shown how semiparametric models can be used for joint normalisation of several time series of data.

List of papers

The thesis is based on the following papers, which will be referred to in the text by their Roman numerals

I. Burdakov O., Grimvall A., Hussian M., and Sysoev O. (2005). Hasse diagrams and the generalized PAV-algorithm for monotonic regression in several explanatory variables. *Submitted to Computational Statistics and Data Analysis*.

II. Hussian M., Grimvall A., Burdakov O., and Sysoev O. (2005). Monotonic regression for the detection of temporal trends in environmental quality data. *MATCH Commun. Math. Comput. Chem.*, **54**, 535-550.

III. Hussian M. and Grimvall A. (2005). Trend analysis of mercury in fish using nonparametric regression. *Department of Mathematics, Division of Statistics*, LIU-MAI-R-2005-07.

IV. Hussian M., Grimvall A., and Petersen W. (2004). Estimation of the human impact on nutrient loads carried by the Elbe River. *Environmental Monitoring and Assessment*, **96**, 15-33.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor ***Anders Grimvall***. He was one big reason that makes this thesis available. His knowledge and experience as a researcher and in supervision is invaluable. I want to thank him for accepting me as one of his students and for his enthusiasm and efforts with me. My studies in Linköping are one experience I will never forget and this is because of him.

My valuable thanks go to my family, my wife ***Mervat***, and my children, ***Hager***, ***Omer*** and ***Toka*** and to the spirit of my ***mother***. I would like to say *thank you for every thing you have done to me*.

Special thanks goes to *Eva Leander, Stig Danielson, Anders Nordgaard, Claudia Libiseller* and *Bodil Stavklint* for there good spirit, for helping and encouraging me to finalise my studies and supporting me in many different aspects.

I want also to thank my colleague and second supervisor *Oleg Burdakov* (optimisation division) for his efforts and being so friendly and helpful in my research and for having such wonderful time during our joint work.

I want also to thank my colleagues, the formal and the new ones at the division of Statistics, Linköping University, who have in one way or another contributed to this work. *Many thanks to all of you*.

Special thanks also go to *Lars Walter* and *Åsa Danielson* who have been like a family to me. They have not forgotten me during my stay in Sweden and they were there when I needed help or needed some one to talk to. For you I would like to send *my love* and *my precious wishes* to you.

Contents

1 Introduction

All data concerning the state of the environment are more or less uncertain. Observational errors can be substantial, measurements can be sparse or based on improper sampling, and the recorded values can be strongly influenced by the weather conditions that prevail before and on sampling occasions. This can make it difficult to determine the causes of observed changes in the environment. In particular, it can be problematic to distinguish between natural variability and the combined effect of a large number of interventions. This thesis is devoted to statistical methods that can be used to reduce irrelevant variation in the collected data and thereby help clarify the impact we humans have on our natural surroundings.

The problems encountered when evaluating trends in time series of environmental quality data can be illustrated by a simple example. Figure 1.1 shows a data set representing observed concentrations of mercury in Atlantic cod caught in the middle of the North Sea (53^{o} 10' N, 2^{o} 5' E). Visual inspection of the collected data gives the impression of a weak downward trend, but the observed concentrations vary considerably.

Closer examination of collected data and other available information revealed that the analysed fish samples represented a wide range of body lengths, and the scatter chart in Figure 1.2 shows that the mercury content usually increased in relation to body length. Consequently, it should be feasible to make more precise statements about mercury trends, if we first remove the variation in measured concentrations that can be attributed to variation in body length.

The key to the mentioned problem is a thorough statistical analysis of the joint distribution of sampling year, body length and mercury content. Figure 1.3 presents a three-dimensional scatter chart of the observed values, and Figure 1.4 depicts a response surface fitted to observed data. Without going into details, it is obvious that the impact of differing body length can be removed or suppressed by investigating how the expected mercury content varies with a fixed body length or a given probability distribution of body lengths. Two questions that must be considered are how response surfaces can best be fitted to observed data and how such surfaces can best be utilised to explain temporal trends in the collected data.

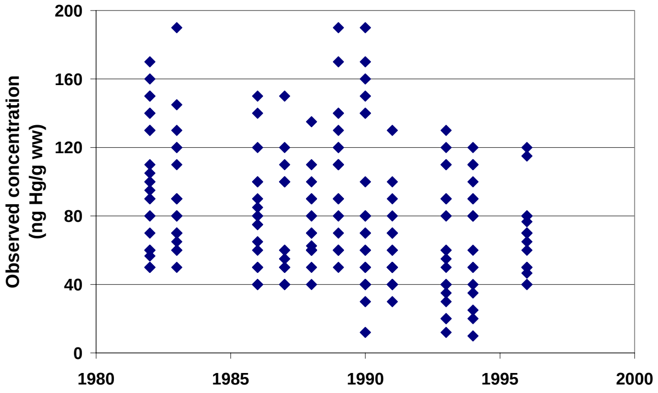


Figure 1.1. Observed concentrations of mercury in muscle tissue from Atlantic cod (Gadu morhua) caught in the North Sea (53^o 10' N, 2^o 5' E).

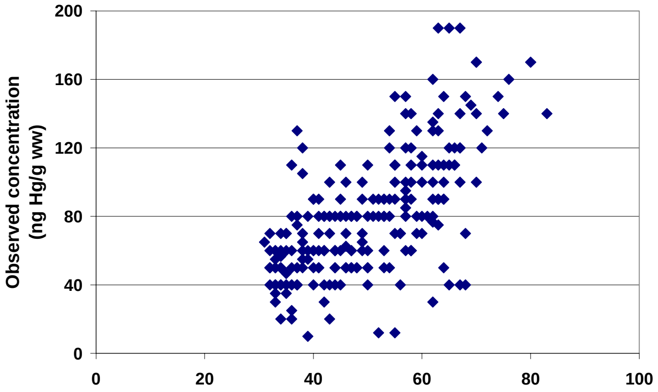


Figure 1.2. Observed concentrations of mercury in muscle tissue from Atlantic cod (Gadu morhua) caught in the North Sea (53^o 10' N, 2^o 5' E) in relation to body length (cm).

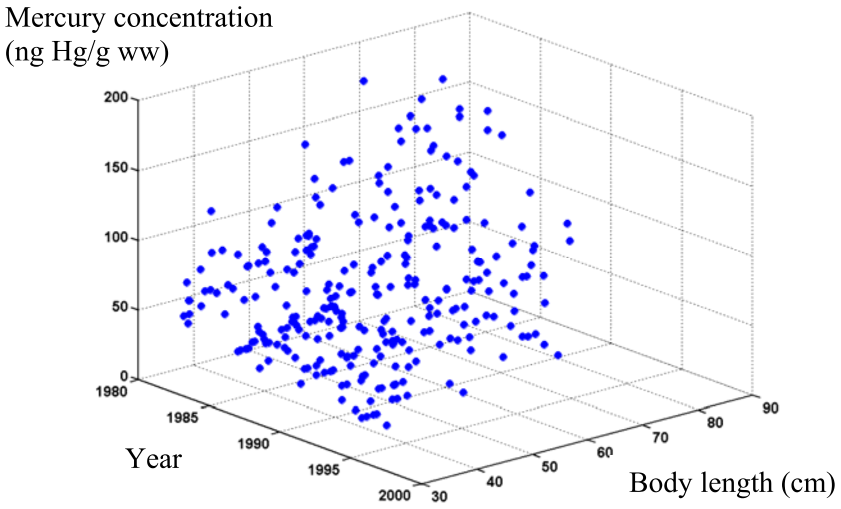


Figure 1.3. Observed concentrations of mercury in muscle tissue from Atlantic cod (Gadu morhua) caught in the North Sea (53^o 10' N, 2^o 5' E).

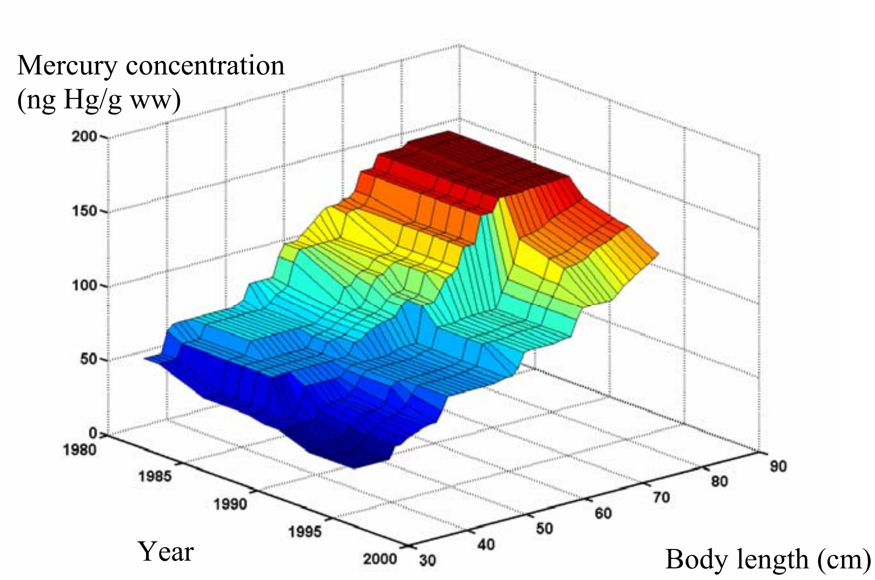


Figure 1.4. Fitted response surface for mercury concentrations in muscle tissue from Atlantic cod (Gadu morhua) caught in the North Sea (53^o 10' N, 2^o 5' E).

1.1 Study objectives

The general goal of the research underlying this thesis was to develop new statistical tools for detecting trends in time series of environmental quality data. Within this frame, we focused on nonparametric and semiparametric regression methods that can be used to clarify the impact of humans on the environment when a substantial fraction of the changes in the collected data over time can be attributed to one or more variables (covariates) representing natural variability. The temporal trends were modelled nonparametrically to enable extraction of nonlinear trends, whereas the adjustment for covariates was based on parametric or nonparametric models.

More specifically, the present studies had the following objectives:

- To examine the performance of algorithms for monotonic regression (MR) that can accommodate two or more explanatory variables (paper I).
- To develop MR-based normalisation techniques for adjustment and trend assessment of environmental quality data that have a monotonic relationship with one or more covariates (paper II).
- To construct regression-based normalisation techniques for time series of environmental quality data collected over several seasons (paper IV).
- To demonstrate the usefulness of the cited methods for assessing trends in substance flows and levels of contaminants in the environment (papers III & IV).

1.2 Outline of the thesis

The rest of this thesis comprises five chapters. In Chapter 2, the concept of normalisation of environmental quality data is introduced, the probabilistic basis for normalisation is explained, and some examples of normalisation formulae are given. In addition, it is shown how semiparametric models can be used for joint normalisation of several time series of data. Chapter 3 is devoted to MR and included examination of the performance of the GPAV (generalised pool adjacent violators) algorithm and the use of MR for environmental applications. Chapter 4 addresses simultaneous normalisation of several time series of data that can be expected to have similar but not necessarily identical trends, and the focus is on how semiparametric (SPR) models can be used to normalise data collected over several seasons. Chapter 5, compares monotonic and semiparametric methods, and also discusses the advantages and disadvantages of the techniques we used. Finally, Chapter 6 lists the major conclusions of the work, along with some issues that require further research.

2 Normalisation of environmental quality data

Various types of statistical adjustments or standardisation methods are widely used to suppress irrelevant variation in collected data. For example, data gathered during several parts of the year are often seasonally adjusted to facilitate comparisons over time. Similarly, incidence rates of diseases are frequently standardised to a given age and sex distribution to facilitate comparison of populations. In environmental science and management, the use of standardisation techniques has long been nearly synonymous with the formation of ratios that are less variable than the original data. However, in the past decade, a number of new methods have been proposed to remove or lessen the effect of irrelevant fluctuations. In this thesis, the term *normalisation* refers to such techniques, and special attention is given to regression-based procedures that aim to clarify the human impact on the environment by removing and suppressing meteorologically driven fluctuations and other natural variation in time series of environmental quality data.

The basic idea of normalisation is simple. If observations of meteorological or other naturally fluctuating variables make us believe that the studied response variable takes a value that is *c* units higher than the mean response, then normalisation implies that we subtract this expected increase *c* from the observed response. A general probabilistic framework for this type of adjustment of collected data has been presented by Grimvall and co-workers (2001), and numerous articles have been published that describe specific normalisation techniques. In particular, several investigators have used regression methods, or related statistical learning techniques such as neural networks, to normalise air quality data (Bloomfield *et al.*, 1996; Holland *et al.*, 1999; Huang & Smith, 1999;

Shively *et al.*, 1999; Gardner & Dorling, 2000*a*,*b*; Thompson *et al.*, 2001). A few authors have considered flow-normalisation of riverine loads of substances (Stålnacke *et al.*, 1999; Stålnacke and Grimvall, 2001; Uhlig & Kuhbier, 2001), and concentrations of pollutants in sediments are often normalised with respect to grain size or amount of organic matter in the analysed samples (e.g., Koelmans *et al.*, 1997; Clark *et al.*, 2000). Additional examples of statistical methods that are used to reduce the variability of observed data include normalisation regarding the fat content in biota or the salinity of water samples collected in estuaries.

2.1 Notation and basic definitions

Here, we discuss normalisation of time series y_t, $t = 1, \dots, n,$ of environmental data quality data that are influenced by random vectors $\boldsymbol{x}_t = (x_{1t}, \dots, x_{pt})$, $t = 1, \dots, n,$ representing the natural forcing of the ecosystem under consideration or other forms of natural variability in the collected data. First, we use the concept of conditional expectation to make a formal definition.

Definition. Let

$$y_t = f(t, \boldsymbol{x}_t) + \varepsilon_t, \quad t = 1, ..., n \tag{2.1}$$

where $\boldsymbol{x}_t$, $t = 1, \dots, n$, are identically distributed random vectors and ε_t, $t = 1, \dots, n$, are independent, identically distributed error terms with mean zero. Further, assume that the error terms are independent of the $\boldsymbol{x}$ vectors. Then, the vector of $\boldsymbol{x}_t$-normalised y values is defined by the equation

$$\tilde{y}_t = y_t - \left(E(y_t \mid \boldsymbol{x}_t) - E(y_t)\right), \ t = 1, ..., n \tag{2.2}$$

where $E(y_t | \boldsymbol{x}_t)$ depicts the conditional expectation of y_t given $\boldsymbol{x}_t$. Moreover, we say that

$$\tilde{y}_t(\boldsymbol{c}) = y_t - \left(E(y_t | \boldsymbol{x}_t) - E(y_t | \boldsymbol{x}_t = \boldsymbol{c})\right), \quad t = 1, \ldots, n \qquad (2.3)$$

is a vector of y values normalised to $\boldsymbol{x} = \boldsymbol{c}$.

The definition of $\boldsymbol{x}$-normalised y values is taken from the above-mentioned report by Grimvall and co-workers (2001), in which the term *global normalisation* was also introduced. The use of y values normalised to $\boldsymbol{x} = \boldsymbol{c}$ is presented in papers II, where the mercury content of fish muscle from Atlantic cod is normalised to a body length of 49.6 cm. Conceptually, the latter type of normalisation is related to the variance reduction that may be achieved by sifting sediments or performing other physical or chemical fractionations of the analysed samples. In contrast to *global normalisation* we shall call it *local normalisation*.

2.2 Normalisation using additive models

The simplest form of normalisation is based on models that have a linear temporal trend and a linear relationship between the observed state of the environment and a set of meteorological covariates or other variables representing natural variability. Let us assume that

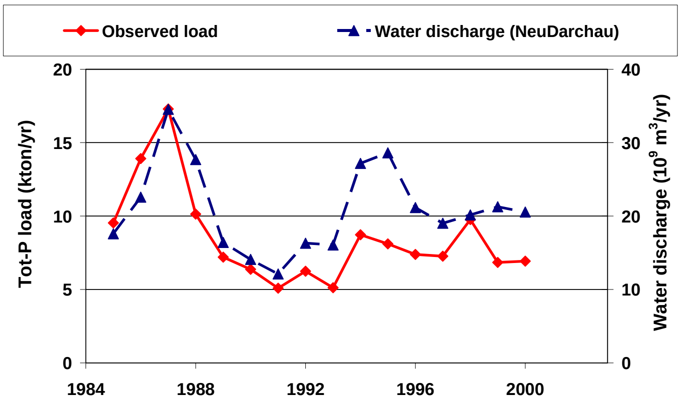
$$y_t = \gamma_0 + \gamma_1 t + \sum_{i=1}^{p} \beta_i x_{it} + \varepsilon_t, \quad t = 1, \ldots, n \qquad (2.4)$$

where β_i, i=1, …, p are regression parameters, and $\gamma_0 + \gamma_1 t$ is a linear trend function. Then, it follows directly from the definition in the previous section that the vector of $\boldsymbol{x}$-normalised y values can be obtained by computing

$$\tilde{y}_t = y_t - \sum_{i=1}^{p} \beta_i (x_{it} - E(x_{it})), \quad t = 1, \dots, n \tag{2.5}$$

Figures 2.1 and 2.2 illustrate the results obtained when this formula was used to normalise annual loads of phosphorus with respect to annual water discharge values, and the model parameters were estimated using ordinary least squares regression.

a)



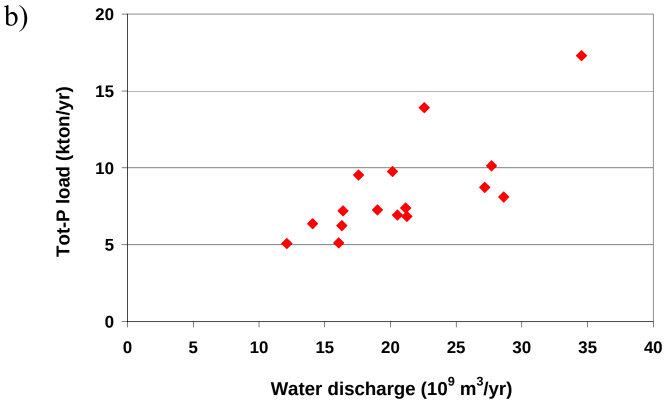


Figure 2.1. Annual loads of total phosphorus (Tot-P) in the Elbe River at Brunsbüttel. The two graphs show the following: time series plots of tot-P loads and water discharge values (a); a scatter chart of tot-P load vs. water discharge (b).

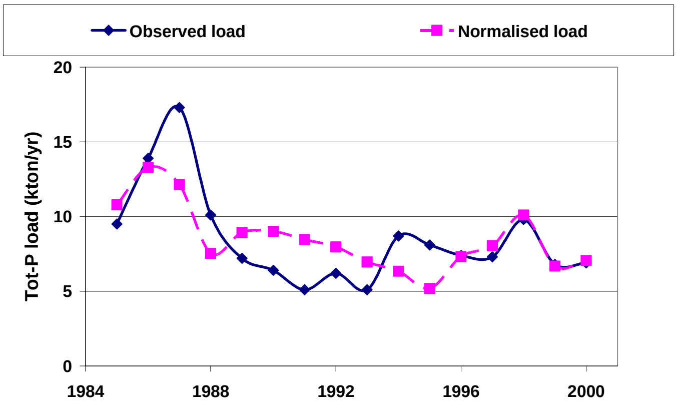


Figure 2.2. Annual loads of total phosphorus (Tot-P) in the Elbe River at Brunsbüttel normalised with respect to water discharge.

The normalisation formulae 2.2 and 2.3 remain unchanged if the linear temporal trend $\gamma_0 + \gamma_1 t$ is replaced with an arbitrary trend function $h(t)$.

Furthermore, it can easily be shown that, for this type of model, global normalisation is identical to local normalisation of y_t to $\boldsymbol{x}_t = E(\boldsymbol{x}_t)$.

The general additive model

$$y_t = h(t) + g(\boldsymbol{x}_t) + \varepsilon_t, \quad t = 1, \ldots, n \tag{2.6}$$

where both h and g may be nonlinear, calls for more thorough examination. First, we note that

$$\tilde{y}_t = y_t - \left(g(\boldsymbol{x}_t) - E(g(\boldsymbol{x}_t))\right), \quad t = 1, \ldots, n \tag{2.7}$$

and

$$\begin{aligned} \tilde{y}_t(\boldsymbol{c}) &= y_t - \left(g(\boldsymbol{x}_t) - E(g(\boldsymbol{x}_t) \mid \boldsymbol{x}_t = \boldsymbol{c})\right) \\ &= y_t - g(\boldsymbol{x}_t) + g(\boldsymbol{c}), \quad t = 1, \ldots, n \end{aligned} \tag{2.8}$$

Moreover, it is simple to show that, for all additive models, local and global normalisation give rise to the same trend slope. This follows directly from the fact that

$$\tilde{y}_t - \tilde{y}_t(\boldsymbol{c}) = \int g(\boldsymbol{c}) dF_{x_t}(\boldsymbol{c}) - g(\boldsymbol{c}), \quad t = 1, \ldots, n \tag{2.9}$$

is constant, if $\boldsymbol{x}_t$ has the same probability distribution for all values of t. Finally, it can be noted that the global normalisation preserves the mean, in other words

$$E(\tilde{y}_t) = E(y_t), \quad t = 1, \ldots, n \tag{2.10}$$

2.3 Normalisation using non-additive models

Varying parameter models form an important class of models that are non-additive in t and $\boldsymbol{x}$. Let us specifically consider models of the form

$$y_t = h(t) + \sum_{i=1}^{p} \beta_{it}\, x_{it} + \varepsilon_t$$

$$= h(t) + \boldsymbol{x_t}\,\boldsymbol{\beta}_t + \varepsilon_t, \quad t = 1, \dots, n \tag{2.11}$$

where $\boldsymbol{\beta_t} = (\beta_{1_t}, \dots, \beta_{p_t})^T$ are time-dependent regression parameters. In this case,

$$\tilde{y}_t = y_t - (\boldsymbol{x}_t - E(\boldsymbol{x}))\boldsymbol{\beta}_t, \quad t = 1, \dots, n \tag{2.12}$$

and

$$\tilde{y}_t(\boldsymbol{c}) = y_t - (\boldsymbol{x}_t - \boldsymbol{c})\boldsymbol{\beta}_t, \quad t = 1, \dots, n \tag{2.13}$$

which implies that global normalisation is identical to local normalisation to $\boldsymbol{x} = E(\boldsymbol{x})$. If $\boldsymbol{c} \neq E(\boldsymbol{x})$, the two normalisation methods can give rise to different trends.

In papers II and III we utilise MR models that are both non-additive and non-linear. Then it is convenient to use the notation

$$y_t = f(t, \boldsymbol{x}_t) + \varepsilon_t, \quad t = 1, \dots, n, \tag{2.14}$$

$$\begin{aligned} \tilde{y}_t &= y_t - \left(E(y_t \,|\, \boldsymbol{x}_t) - E(y_t)\right) \\ &= y_t - f(t, \boldsymbol{x}_t) + \int f(t, \boldsymbol{c})\, dF_{X_t}(\boldsymbol{c}), \quad t = 1, \dots, n \end{aligned} \tag{2.15}$$

$$\begin{aligned}\tilde{y}_t(\boldsymbol{c}) &= y_t - \left(E(y_t \mid \boldsymbol{x}_t = \boldsymbol{c}) - E(y_t)\right) \\ &= y_t - f(t, \boldsymbol{x}_t) + f(t, \boldsymbol{c}), \quad t = 1, ..., n\end{aligned} \tag{2.16}$$

and it can be noted that the difference between global and local normalisation

$$\tilde{y}_t - \tilde{y}_t(\boldsymbol{c}) = \int f(t, \boldsymbol{c}) dF_{x_t}(\boldsymbol{c}) - f(t, \boldsymbol{c}), \quad t = 1, ..., n \tag{2.17}$$

may vary over time, regardless of how $\boldsymbol{c}$ is selected.

2.4 Simultaneous normalisation of several time series of data

In paper IV, we normalised time series of riverine loads of nitrogen and phosphorus. The model used is a generalisation from one to several explanatory variables of the SPR model proposed by Stålnacke and Grimvall (2001). To be more precise, we assume that the relationship between the riverine load y_{tj} for the *j*th month of the *t*th year and the contemporaneous values of *p* explanatory variables $x_{1t}^{(j)}, ..., x_{pt}^{(j)}$ has the form

$$y_t^{(j)} = \alpha_t^{(j)} + \sum_{i=1}^{p} \beta_i^{(j)} x_{it}^{(j)} + \varepsilon_t^{(j)}, \quad t = 1, \dots, n, \quad j = 1, ..., m \tag{2.18}$$

where $\alpha_t^{(j)}$, *t* = 1, …, *n*, *j* = 1, …, *m*, represent deterministic trends, and the error terms are independent of each other and of the explanatory variables. Recently, Libiseller and Grimvall (2005) and Giannitrapani *et al.* (2005) addressed similar normalisation problems when they examined air quality and deposition data representing several wind sectors. Here, we emphasise that both the riverine loads by season and the deposition by wind sector can be regarded as multivariate time series of data. Similarly, data representing several sampling sites along a gradient or several congeners of an organic pollutant can be regarded as multivariate data for which a joint analysis would be desirable.

In principle, there is no relationship between the different time series of data in formula 2.18, because all error terms are assumed to be statistically independent and each series has its own set of intercept and slope parameters. However, when the model parameters are estimated, it is natural to introduce constraints on the variability of $\alpha_t^{(j)}$ and $\beta_i^{(j)}$, j=1,…,m across the different time series of data. This is discussed further in the next section.

2.5 Estimation of normalisation models

We have already noticed that the simple linear normalisation model can be estimated using ordinary least squares regression. Estimation of semiparametric normalisation models such as 2.18 is, however, a more intricate task. First, we note that the number of parameters is larger than the number of observations. In a non- or semiparametric setting, this requires that smoothness conditions be introduced to decrease the degrees of freedom and render the model estimable. Second, we see that smoothness conditions for the intercept parameters are introduced in a more natural manner, if we reformulate the model so that the intercept represents the expected response when the covariates are equal to their expectations, as follows:

$$y_t^{(j)} = \alpha_t^{(j)} + \sum_{i=1}^{p} \beta_i^{(j)} (x_{it}^{(j)} - E(x_{it}^{(j)})) + \varepsilon_t^{(j)}, \quad t = 1, \dots, n, \quad j = 1, \dots, m \quad (2.19)$$

In the calculations described in paper IV, a roughness penalty approach is used to SPR, that is, the parameters are estimated by minimising an objective function that has two components: the residual sum of squares, and a measure of the roughness of $E(Y_t^{(j)})$ regarded as a function of *t* and *j* (Green and Silverman, 1994).

Compared to other smoothing methods, such as kernel smoothing and splines (Härdle, 1997; Hastie *et al.,* 2001; Schimek, 2001), an advantage of the roughness penalty approach is that the smoothness conditions can easily be adapted to the type of data analysed. For example, when data are collected over several seasons, it would be natural to introduce constraints that force the intercept to vary smoothly with both year and season. Furthermore, it is natural to claim that the trend levels for the last season in one year and the first season in the following year are close to each other. We use the term spiral smoothing for this type of constraints on the intercept parameters (see Figure 2.3a). For annual data representing different wind sectors it is more appropriate to employ some form of circular smoothing (Figure 2.3b), whereas data collected along a gradient may require smoothing in two directions: over time and along the sampled gradient (Figure 2.3c). Paper IV and Chapter 4 give further details about roughness penalty approaches to the estimation of SPR models.

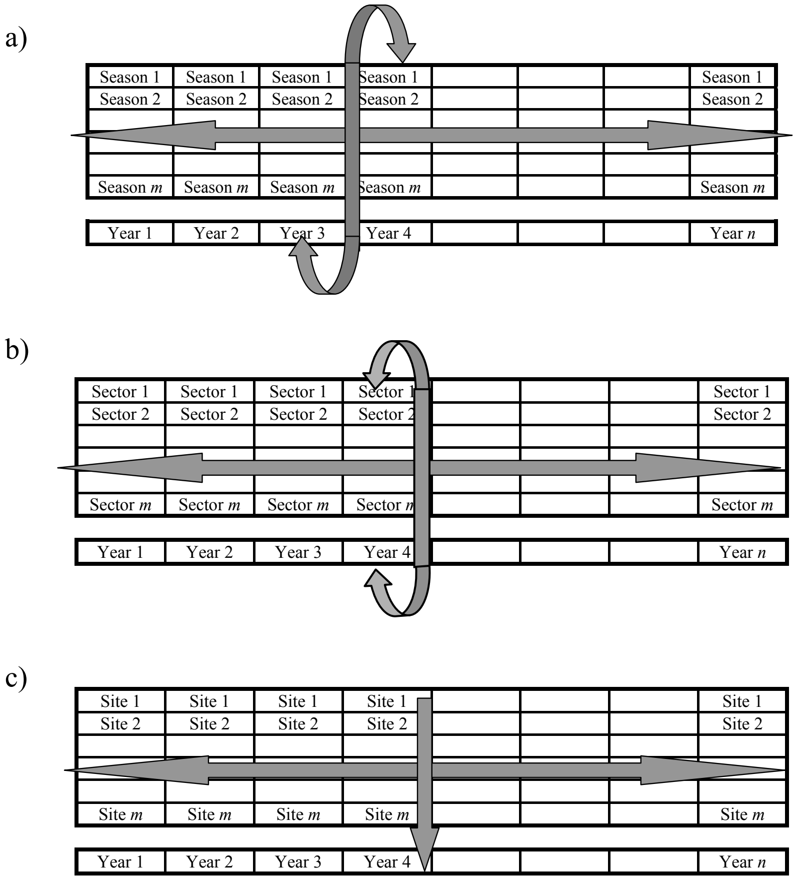


Figure 2.3. Different types of smoothing patterns for the intercept of the SPR model 2.4. The three graphs show spiral smoothing for data collected over several season (a), circular smoothing for data representing several sectors (b), and gradient smoothing for data collected along a gradient (c).

Estimation of MR models for normalisation requires a two-step procedure. Algorithms for such regression provide fitted values

$$\hat{y}_t = \hat{f}(t, \boldsymbol{x}_t), \quad t = 1, ..., n \tag{2.20}$$

for those $\boldsymbol{x}$ vectors that are linked to an observed response value. However, the normalisation may require estimates of the expected response at additional points. For instance, in the case of local normalisation, we might want to estimate $\tilde{y}_t(\boldsymbol{c})$ by computing

$$y_t - \hat{f}(t, \boldsymbol{x}_t) + \hat{f}(t, \boldsymbol{c}), \quad t = 1, ..., n \tag{2.21}$$

where $\boldsymbol{c}$ is not necessarily identical to any of the $\boldsymbol{x}_t$ values. Likewise, in global normalisation, it must be possible to compute

$$y_t - \hat{f}(t, \boldsymbol{x}_t) + \int \hat{f}(t, \boldsymbol{c})\, dF_{X_t}(\boldsymbol{c}), \quad t = 1, ..., n \tag{2.22}$$

or

$$y_t - \hat{f}(t, \boldsymbol{x}_t) + \frac{1}{n}\sum_{j=1}^{n} \hat{f}(t, \boldsymbol{x}_j), \quad t = 1, ..., n \tag{2.23}$$

The specific algorithms that are needed to obtain the fitted values

$$\hat{y}_t = \hat{f}(t, \boldsymbol{x}_t), \quad t = 1, ..., n \tag{2.24}$$

are discussed extensively in paper I and Chapter 3, whereas the extrapolation of $\hat{f}$ to new $\boldsymbol{x}$ values is briefly addressed in papers II and III.

2.6 Further notes on normalisation

The theoretical definition of normalisation that was given in section 2.1 can easily be extended to accommodate serially dependent error terms. Whether or not it is feasible to estimate such models is determined by the function $f(t, \boldsymbol{x}_t)$. Many semiparametric normalisation models are estimated using back-fitting algorithms in which parametric and nonparametric subroutines alternate. In such cases, it may be sufficient to replace an ordinary least squares estimator with a maximum-likelihood estimator that can accommodate serially dependent error terms. In contrast, the MR algorithms can only be applied to models that have independent error terms.

The stationarity of the sequence $\boldsymbol{x}_t$, $t = 1, \ldots, n$, is crucial for global normalisation. The idea of using that type of normalisation to remove natural fluctuations in the collected data without distorting the anthropogenic trend is based on the assumption that the probability distribution of $\boldsymbol{x}_t$ is constant in time. Local normalisation is less demanding in this respect, because $\boldsymbol{x}$ is kept fixed. However, any normalisation with respect to a variable that has a time-dependent distribution will raise questions regarding spurious trends and the risk of concealing the effect that humans have on the environment.

Finally, it should be emphasised that none of the normalisation models discussed in this thesis include explicit information about the anthropogenic forcing of the ecosystem under consideration. Nevertheless, such models can be constructed. The basic definitions needed have been given by Grimvall *et al.* (2001). Examples illustrating this approach have been published by Forsman and Grimvall (2003) and Wahlin *et al.* (2004) who used physics-based hydrological models to examine meteorologically normalised model outputs.

3 Monotonic regression (MR) models

Monotonic relationships occur in a great variety of contexts. For example, dose-response curves in toxicological or medical experiments are very often monotonic and S-shaped. In environmental systems, the intensity or magnitude of the investigated phenomena is commonly a monotonic function of both the anthropogenic forcing and naturally varying factors, such as temperature, wind speed, and rainfall. Moreover, it is well known that contaminant levels in environmental samples can increase with the age or size of the analysed organism or the content of fat or organic matter in the analysed sample. This necessitates statistical methods that can estimate models in which the expected response increases or decreases in relation to one or more explanatory variables.

MR is a nonparametric method for estimation of models that can be written

$$y_j = f(x_{1j}, ..., x_{pj}) + \varepsilon_j, \quad j = 1, ..., n \tag{3.1}$$

where y is the response variable, f is increasing or decreasing in each of the p explanatory x-variables, and the error terms $\{\varepsilon_j\}$ are independent of each other and of the x-variables. These models have been investigated since the 1950s (Ayer *et al.*, 1955; Robertson and Waltman, 1968; Barlow *et al.*, 1972; De Simone *et al.*, 2001), and they retain the monotonicity of linear models, but relax the strong assumption of linearity (Schell and Singh, 1997). In particular, they are able to capture both nonlinear and non-additive responses to an arbitrary set of covariates. The special case when the expected response is increasing in all explanatory variables is referred to as the isotonic regression (IR). Unless otherwise stated, we shall restrict ourselves to IR, because a simple change of sign of some of the explanatory variables can transform any MR to an IR.

3.1 MR regarded as an optimisation problem

The estimation of MR/IR models can be formulated as an optimisation problem in which a loss function is minimised under a set of simple constraints (Ayer *et al.*, 1955; Robertson and Waltman, 1968; Barlow *et al.*, 1972; De Simone *et al.*, 2001).

Let

$$\begin{array}{cccccc} x_{11} & . & . & . & x_{p1} & y_1 \\ x_{12} & . & . & . & x_{p2} & y_2 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{1n} & . & . & . & x_{pn} & y_n \end{array} \qquad (3.2)$$

denote *n* observations of *p* explanatory variables $x_1, \ldots, x_p$ and one response variable *y*. Then, we can estimate the isotonic relationship between the expected response *E*(*y*) and the covariates $x_1, \ldots, x_p$ by computing fitted values $z_i = \hat{f}(x_{1i}, ..., x_{pi})$ that minimise

$$\sum_{i=1}^{n}(y_i - z_i)^2 \qquad (3.3)$$

under the constraints

$$z_i \le z_j \text{, if } x_{ki} \le x_{kj} \text{, for all } k = 1, ..., p \qquad (3.4)$$

By introducing the partial order

$$\boldsymbol{x}_i \prec \boldsymbol{x}_j \text{ if and only if } x_{ki} \le x_{kj}, \text{ for all } k=1, ..., p$$

on the set of vectors $\{\boldsymbol{x}_i = (\mathrm{x}_{1i}, \ldots, x_{pi}), i = 1, \ldots, n\}$, we can also write the constraints in (3.4) as

$$z_i < z_j \text{ if } \boldsymbol{x}_i \prec \boldsymbol{x}_j \tag{3.5}$$

In other words, IR transfers the partial order on the set of $\boldsymbol{x}$ vectors to the fitted response values.

3.2 The PAV and GPAV algorithms

When $p = 1$, it is easy to compute the least squares solution to the IR problem. The best known algorithm to handle this problem is the PAV (Pool Adjacent Violators) algorithm (Ayer *et al.*, 1955; Barlow *et al.*, 1972; De Simone *et al.*, 2001), for which the point of departure is a data set $M_n = \{(x_i, y_i), i = 1,\ldots, n\}$ that is sorted so that $x_1, \ldots, x_n$ form a non-decreasing sequence. If $n = 1$, it is obvious that the optimal solution is $z_1 = y_1$. If $n = 2$, it is evident that we should set $z_1 = y_1$ and $\mathrm{z}_2 = y_2$ if $y_1 \leq y_2$, or otherwise pool the two observed values and set $z_1 = z_2 = (y_1 + y_2)/2$. The optimal solution for an arbitrary data set M_n can be obtained by recursively solving the IR problem for the data sets M_k, $k = 1, \ldots, n$. To be more precise, we extend the optimal solution for M_k to a preliminary solution for M_{k+1} by setting $z_{k+1} = y_{k+1}$, and then we remove possible violations of the monotonicity by pooling adjacent z values in a backward movement from right to left. Figure 3.1 shows a scatter plot of observed values y and the clusters of fitted values that form the solution produced by the PAV algorithm.

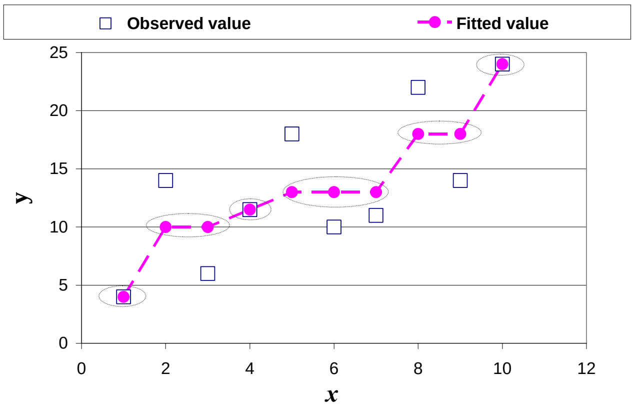


Figure 3.1. Observed and fitted response values using the PAV algorithm. The final solution consists of six clusters of identical values.

If $p > 1$, it is less obvious how the IR problem should be solved. A simple back-fitting procedure based on the PAV algorithm can handle relatively small data sets in which the explanatory variables vary over only a few levels (Dykstra and Robertson, 1982; Bril *et al.*, 1984; Salanti and Ulm, 2001). Moreover, there are adequate algorithms for partial orders that have specific structures, for instance, tree or star structures (Pardalos and Xue, 1999). Other methods, such as simple averaging techniques, are more generally applicable (Mukarjee, 1988; Mukarjee and Stern, 1994; Strand, 2003), but the solutions obtained can be rather far from optimal in the sense of least squares. There are also other techniques that provide accurate solutions but are computationally very expensive for large data sets (Best and Chakravarti, 1990).

This thesis examines the performance of a recently published generalisation of the PAV algorithm from completely to partially ordered data (Burdakov *et al.*, 2004; 2005). This generalisation is referred to as the GPAV (Generalised Pool Adjacent Violators) algorithm, and it resembles the ordinary PAV algorithm in several respects. First, it is recursive in that a solution for the data set $M_n = \{(x_{1i}, ..., x_{pi}, y_i), i = 1, ..., n\}$ is obtained by starting from a solution for M_1, which is then modified into a solution for M_2, and so on. Second, the data are presorted so that the case ($\boldsymbol{x}_i$, y_i) is entered into the calculations before ($\boldsymbol{x}_j$, y_j), if the two $\boldsymbol{x}$ vectors are distinct and $\boldsymbol{x}_i \prec \boldsymbol{x}_j$. Third, monotonicity violators are removed by forming clusters of adjacent cases, and letting the fitted values in each cluster be identical and equal to the observed mean response in that cluster. However, there is also an important difference between the GPAV solutions for $p > 1$ and the PAV/GPAV solutions for $p = 1$. When $p > 1$, the solutions may depend on the order in which the $\boldsymbol{x}$ vectors are entered into the calculations, and many different orderings may be consistent with the given partial order.

The cited articles by Burdakov and co-workers show that the GPAV algorithm normally produces optimal or close to optimal solutions. Furthermore, it is demonstrated that this algorithm has complexity $O(n^2)$, where n is the number of observations. The following discussion is focused on presorting of the data and on some statistical aspects of obtained solutions.

3.3 Hasse diagrams and MR

Figure 3.2 shows an example of a partially ordered set of elements in the Euclidean space R^2 along with a Hasse diagram (Davey and Priestly, 2002) describing this partial order. To make the diagram as simple as possible, edges that are implied by transitivity have been omitted. Furthermore, it can be seen that the vertical positions of the elements define a grouping that is consistent with the given partial order, i.e. $\boldsymbol{x}_j$ is assigned a higher level than $\boldsymbol{x}_i$ if $\boldsymbol{x}_i < \boldsymbol{x}_j$. We used this grouping into levels to topologically sort the elements entered into the GPAV algorithm. More specifically, we used the levels defined by two slightly different Hasse diagrams obtained by ordering all elements from the bottom and top, respectively (see paper I).

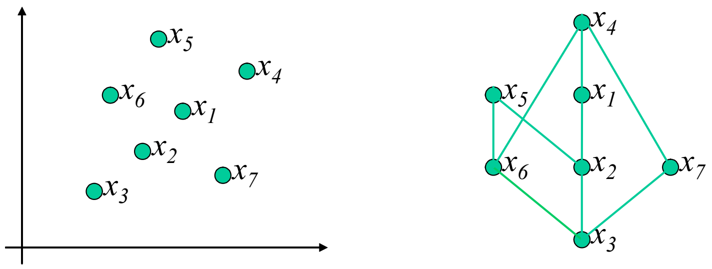


Figure 3.2. Example of a partially ordered set of elements in the Euclidean space R^2 and a Hasse diagram of the partial order.

We conducted a simulation study to examine how the performance of the GPAV algorithm was influenced when data were ordered according to their level in a Hasse diagram (paper I). This was achieved by comparing the following presorting methods:

GPAV-R. The original, randomly ordered data were sorted using a quick-sort algorithm.

GPAV-H1. The output of the quick-sort algorithm was sorted according to the levels of a Hasse diagram drawn from bottom to top.

GPAV-H2. The output of the quick-sort algorithm was sorted according to the levels of a Hasse diagram drawn from top to bottom.

Data sets were generated using the equation

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, ..., n \tag{3.6}$$

where the values of the explanatory variables were drawn from a bivariate normal distribution with mean zero and covariance matrix

$$\boldsymbol{C} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{3.7}$$

and the error terms $\{\varepsilon_j\}$ were independent and identically distributed. Light- or heavy-tailed distributions of the error terms were generated according to normal and double exponential distributions with mean zero and variance one.

Table 3.1 shows the goodness-of-fit defined as

$$\sum_i (\hat{y}_i - y_i)^2 / n \tag{3.8}$$

and the accuracy defined as

$$\sum_i (\hat{y}_i - E(y_i \mid \boldsymbol{x}_i))^2 / n \tag{3.9}$$

when the regression parameters β_1 and β_2 were set to one. Apparently, the presorting with a Hasse diagram significantly improved both the goodness-of-fit and the accuracy of the obtained solutions, especially for the largest samples (n = 10,000). Also, the results indicate that the GPAV-H algorithms, in contrast to GPAV-R, produce consistent estimates of the expected response.

Table 3.2 shows the computational burden involved in MR when the GPAV algorithm was implemented in Visual Basic for Excel and the code was written to minimise the need for memory capacity (paper I). It can be seen that data sets comprising 10,000 observations are easy to handle. Furthermore, it is obvious that, computationally, the most expensive part of the proposed algorithms is the listing of non-redundant constraints. Also, it is worth noting that additional explanatory variables can be handled without any problems, because they enter the calculations only through the partial order of the $\boldsymbol{x}$ vectors.

Table 3.1. Goodness-of-fit and accuracy of the GPAV solutions obtained when the regression parameters β_1 and β_2 were set to one, and different methods were used to presort the data entered into the GPAV algorithm. The number of simulated data sets was 1000 for the two smaller sample sizes (n = 100 and 1000) and 100 for the largest sample size (n = 10000). The values given within parentheses represent standard errors of the estimated means.

Model	n	Goodness-of-fit			Accuracy		
		GPAV-R	GPAV-H1	GPAV-H2	GPAV-R	GPAV-H1	GPAV-H2
$\rho = 0$ Normal errors	100	0.434 (0.003)	0.412 (0.003)	0.411 (0.003)	0.360 (0.002)	0.344 (0.002)	0.343 (0.002)
	1000	0.952 (0.002)	0.724 (0.001)	0.722 (0.001)	0.280 (0.001)	0.109 (0.0003)	0.108 (0.0003)
	10000	1.526 (0.003)	0.906 (0.001)	0.906 (0.001)	0.600 (0.002)	0.039 (0.0003)	0.040 (0.0003)
$\rho = 0$ Double exp. errors	100	0.456 (0.005)	0.432 (0.005)	0.431 (0.005)	0.344 (0.003)	0.327 (0.003)	0.325 (0.003)
	1000	0.973 (0.003)	0.739 (0.002)	0.736 (0.002)	0.283 (0.001)	0.104 (0.0004)	0.103 (0.0004)
	10000	1.565 (0.039)	0.905 (0.021)	0.905 (0.021)	0.636 (0.029)	0.037 (0.003)	0.037 (0.003)
$\rho = 0.9$ Normal errors	100	0.542 (0.003)	0.531 (0.003)	0.529 (0.003)	0.236 (0.002)	0.230 (0.002)	0.229 (0.002)
	1000	0.874 (0.001)	0.797 (0.001)	0.795 (0.001)	0.110 (0.0003)	0.064 (0.0002)	0.064 (0.0002)
	10000	1.090 (0.001)	0.929 (0.001)	0.929 (0.001)	0.145 (0.0003)	0.020 (0.0001)	0.020 (0.0001)
$\rho = 0.9$ Double exp. errors	100	0.556 (0.005)	0.547 (0.005)	0.546 (0.005)	0.224 (0.002)	0.218 (0.002)	0.217 (0.002)
	1000	0.875 (0.002)	0.807 (0.002)	0.806 (0.002)	0.101 (0.0003)	0.062 (0.0003)	0.061 (0.0003)
	10000	1.082 (0.002)	0.927 (0.002)	0.928 (0.002)	0.140 (0.0004)	0.020 (0.0001)	0.020 (0.0002)

Table 3.2. Average CPU-time for different parts of the GPAV approach to IR when a memory-conserving algorithm was implemented in Visual Basic for Excel. The calculations were performed on a PC (1.5 GHz) running under Windows XP.

No. of observations	Average CPU time (s)				
	Quicksorting observed data	Listing non-redundant constraints	Running GPAV-R	Running GPAV-H1 and H2	Total time
100	0.001	0.013	0.005	0.013	≈ 0.03
1000	0.012	1.61	0.41	0.76	≈ 3
10000	0.15	198	36	76	≈ 300

In conclusion, the use of Hasse diagrams to presort observed data greatly improves the performance of the GPAV algorithm, and the current implementations of GPAV enable convenient analysis of fairly large data sets. Also, the very general form of the model makes it an attractive choice in many applications. However, this flexibility has a price. Closer examination of the results in Table 3.1 reveals that the mean square residual can be substantially smaller than the true variance of the error terms, that is, the good fit to observed data can be partly attributed to over-fitting. It has been suggested that the number of clusters in the obtained solution can be interpreted as the degrees of freedom of the model (Mallows, 1973; Mallows, 1995; Shell and Singh, 1997; Meyer and Woodroofe, 2000), and, in principle, this could provide a suitable adjustment of the mean square residual. However, our simulations reported in paper I demonstrated that, even after the indicated adjustment, the true error variance is still underestimated.

3.4 Estimation of monotonic response surfaces for environmental data

The use of MR to estimate monotonic response surfaces can be illustrated with two examples. First, we can consider a study of temporal trends in the concentration of total nitrogen in the Stockholm archipelago (Libiseller *et. al.*, 2005). Due to the random mixing of fresh water and sea water, there was a substantial variability in the collected water quality data. Furthermore, it was difficult to get an overview of all data because they represented several different depths at several stations. However, MR of the total nitrogen concentration on both time and salinity provided a useful summary of all data, and, after extrapolating the fitted response values to a (monotonic) response surface (papers II and III) we obtained the graphs shown in Figures 3.3 and 3.4. It is particularly noticeable that, in less saline waters, there was a clear response to the introduction of improved wastewater treatment in 1995. The trend at higher salinity levels was much weaker, possibly nonexistent.

Figure 3.5 illustrates a set of monthly flow-weighted concentrations of total nitrogen in the Elbe River at Brunsbüttel (downstream of Hamburg) in Germany. Upon initial inspection of the diagram, it may seem impossible to apply MR for such data. However, the seasonal pattern could be decomposed into one non-decreasing and one non-increasing phase (see paper II), which enabled a joint analysis of the trend and seasonal components. Figure 3.6 shows the results obtained when the seasonal component was assumed to have a maximum in March and a minimum in August.

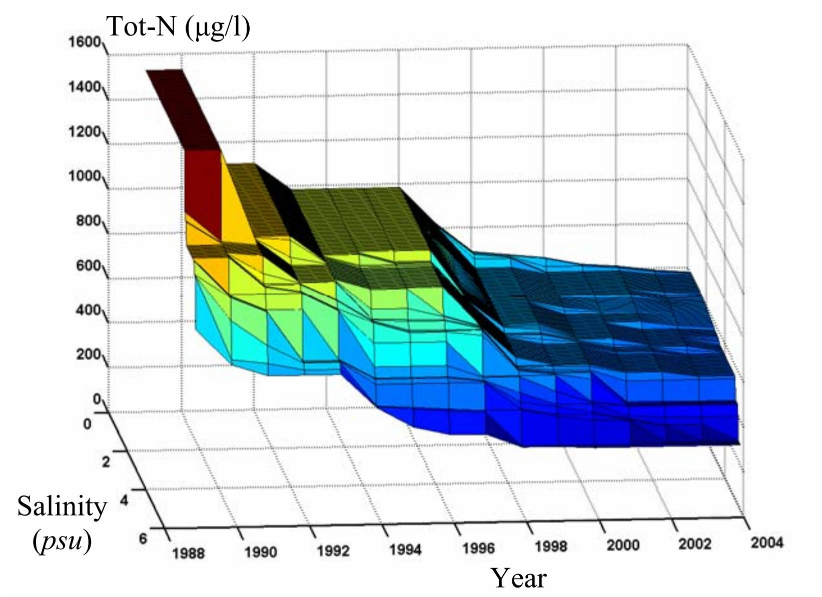


Figure 3.3. Fitted response surface for all nitrogen and salinity data collected in September at the stations in the inner part of Stockholm archipelago.

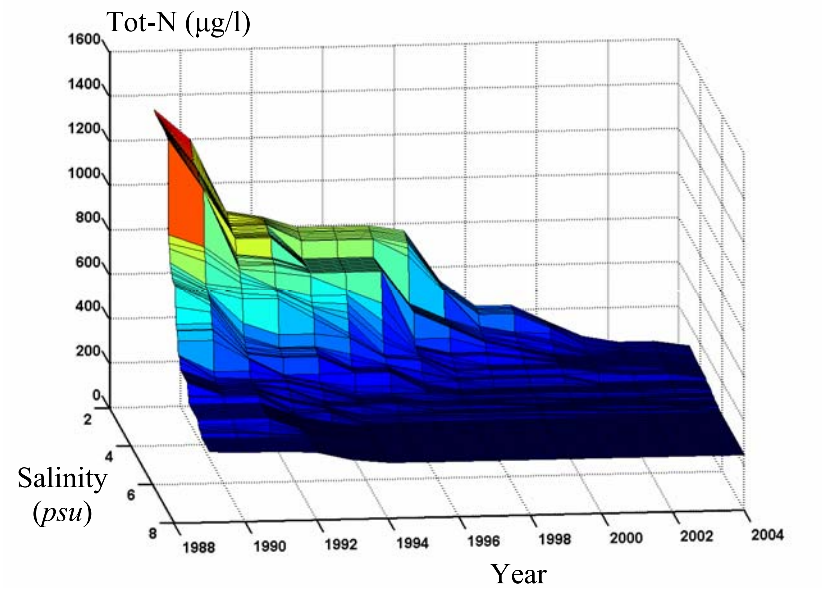


Figure 3.4. Fitted response surface for all nitrogen and salinity data collected in September at the stations in the outer part of Stockholm archipelago.

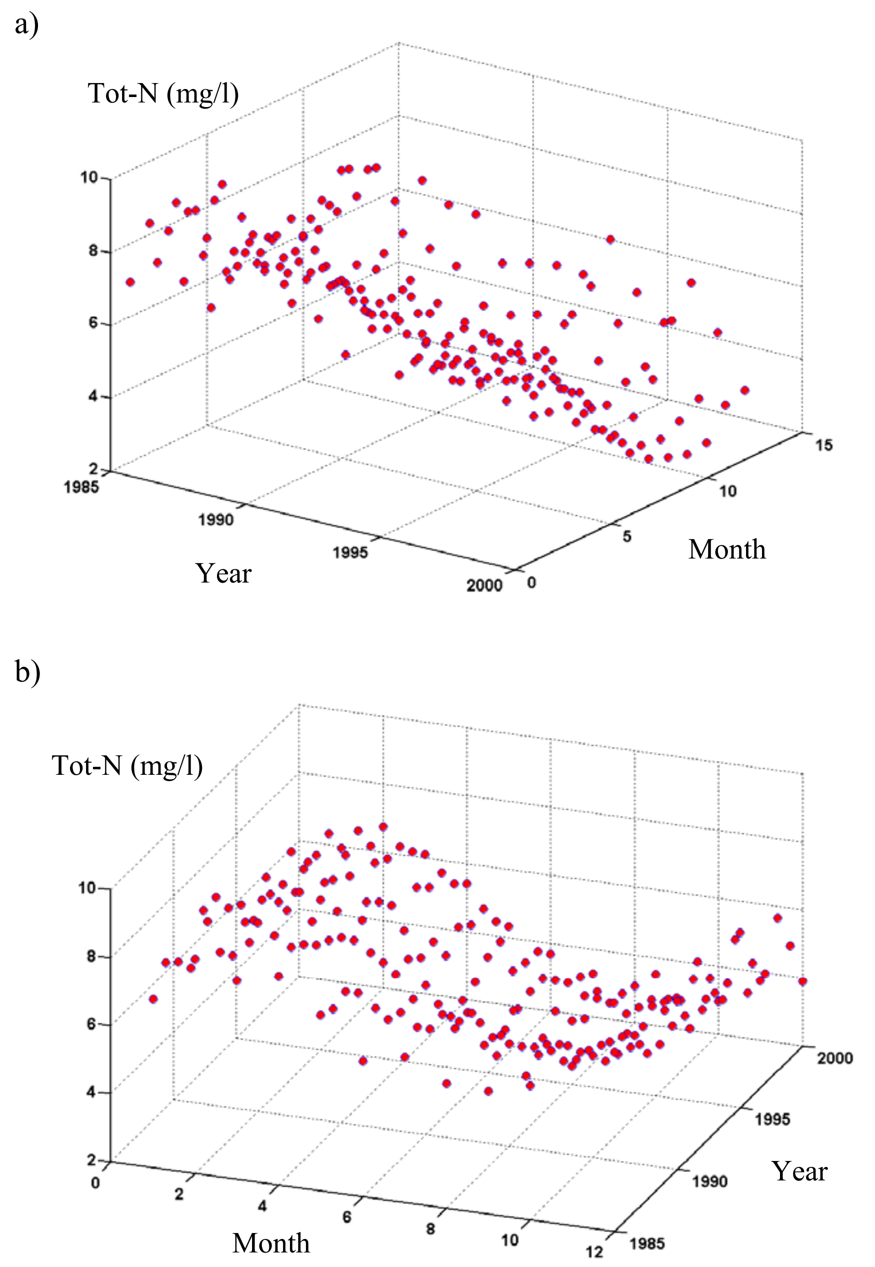


Figure 3.5. Monthly mean concentrations of total nitrogen (Tot-N) measured in the Elbe River at Brunsbüttel downstream of Hamburg City, Germany.

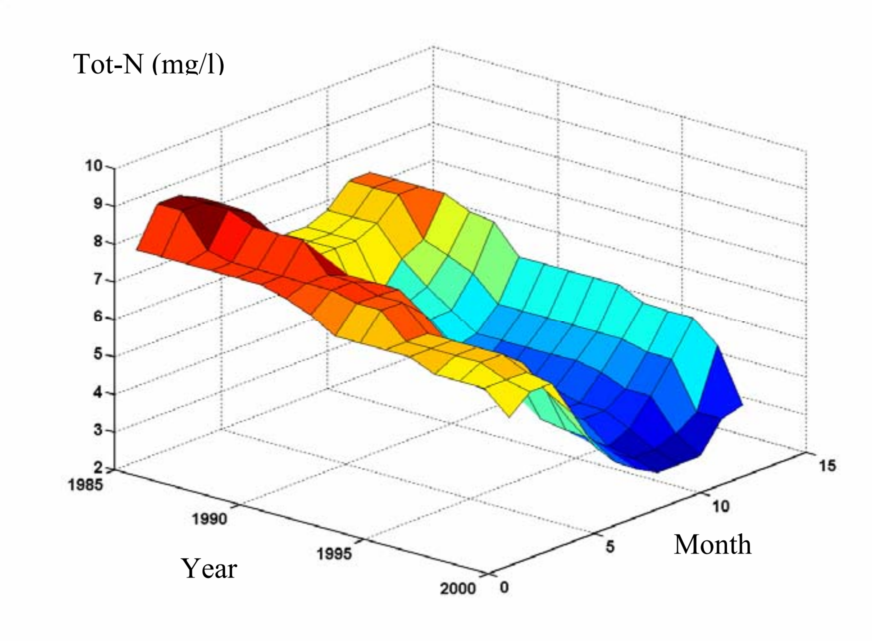


Figure 3.6. Response surface obtained by applying MR to monthly mean concentrations of total nitrogen (Tot-N) in the Elbe River at Brunsbüttel.

3.5 Normalisation using MR

To illustrate the MR-based normalisation methods discussed in Chapter 2, the GPAV algorithm is employed to compute fitted response values associated with the observed ***x*** vectors. Thereafter, an extrapolation method that preserves monotonicity is used to extrapolate (or interpolate) these fitted values to a response surface. Finally, the obtained response surface is used to normalise all observed values with respect to the covariates under consideration.

When this procedure was employed to normalise the data in Figure 1.1 with respect to body length, we obtained the results illustrated in Figure 3.7.

Apparently, the normalisation removed a substantial part of the variability of the collected data. In addition, it can be noted that the downward tendency that may be discerned in the original data is absent in the normalised data.

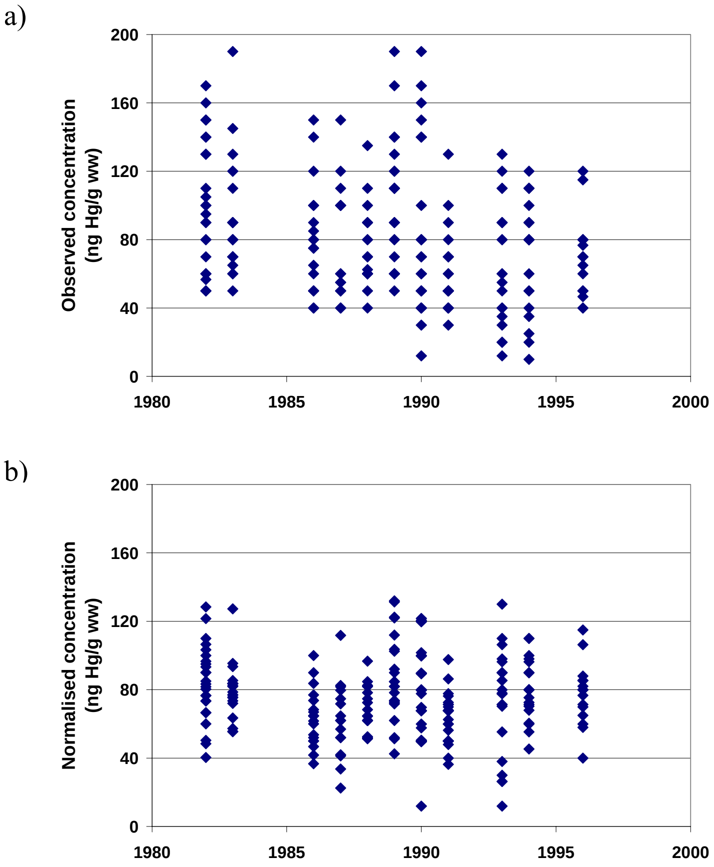


Figure 3.7. Concentrations of mercury in muscle tissue from Atlantic cod (Gadu morhua) caught in the North Sea (53^o 10' N, 2^o 5' E). The illustrated results represent observed concentrations (a), and data normalised to a body length of 49.6 cm (b).

4 Semiparametric regression (SPR) models

The deterioration of an ecosystem is usually a slow process and the change from one year to the next can be difficult to discern even if the long-term trend is severe. This situation requires statistical models in which the anthropogenic impact is modelled as a smooth function of time. In Chapter 2, we introduced two types of SPR models that facilitate simultaneous extraction of smooth, possibly nonlinear, trends and adjustment for covariates. The first of these models (2.11) was intended for assessment of trends in a single response variable, whereas the second (2.18) was designed to enable joint assessment of trends in several time series of data.

Models in which the trend function is specified in a nonparametric fashion can be justified by a desire to make unprejudiced inference about the shape of the trend curve. Theoretically, similar arguments could rationalise the use of models in which both the trend and the influence of covariates are specified nonparametrically. However, models with very few structural constraints may lead to problems with over-fitting, unless the roughness of the estimated response surfaces is penalised when the model is estimated. This calls for statistical procedures in which the degree of smoothness is selected just as carefully as other model features. This chapter will show that the advantage of the above-mentioned semiparametric models is that cross-validation can be applied to combine a roughness penalty approach with a flexible selection of smoothing parameters.

4.1 Estimation of SPR models

Let us now consider semiparametric models of the form

$$y_t^{(j)} = \alpha_t^{(j)} + \sum_{i=1}^{p} \beta_i^{(j)} x_{it}^{(j)} + \varepsilon_t^{(j)}, \quad t = 1, ..., n, \quad j = 1, ..., m \qquad (4.1)$$

where $y_t^{(j)}$ is the observed response for the *j*th season of the *t*th year, $x_{it}^{(j)}$, *i* = 1, …, *p* represent contemporaneous values of *p* explanatory variables standardised to mean zero and variance one, and $\varepsilon_t^{(j)}$ are independent identically distributed random error terms that have mean zero and are independent of the explanatory variables. As can be seen, the slope parameters ($\beta_i^{(j)}$, *i* = 1, …, *p*) in this model are allowed to vary with the season under consideration, whereas the intercept $(\alpha_t^{(j)}, t = 1, .., n,$ $j = 1, ..., m)$ is permitted to vary nonparametrically with both the year and the season.

Following a general procedure outlined by Green and Silverman (1994), and implemented by Stålnacke and Grimvall (2001) for models with a single covariate (*p* = 1), we introduced the penalised sum of squared residuals

$$\begin{aligned} S(\alpha, \beta) = & \sum_{t,j} (y_t^{(j)} - \alpha_t^{(j)} - \beta_1^{(j)} x_{1t}^{(j)} - ... - \beta_p^{(j)} x_{pt}^{(j)})^2 \\ & + \lambda_1 \sum_{t,j} (\alpha_t^{(j)} - \frac{\alpha_{t+1}^{(j)} + \alpha_{t-1}^{(j)}}{2})^2 + \lambda_2 \sum_{t,j} (\alpha_t^{(j)} - \frac{\alpha_t^{(j-1)} + \alpha_t^{(j+1)}}{2})^2 \end{aligned} \qquad (4.2)$$

where the first roughness penalty factor (λ_1) controls the interannual variation of the intercept parameters, and the second such factor (λ_2) controls the variation over seasons for these parameters. For fixed penalty factors, the model parameters were estimated using a back-fitting algorithm in which nonparametric estimation of the intercept is alternated with ordinary least squares estimation of the slope parameters (paper IV).

Schimek (2001) has advocated that back-fitting is less reliable than joint estimation of all model parameters. However, extensive application of our back-fitting algorithm to time series of environmental data has not been associated with any convergence problems, and the computational burden is moderate.

The roughness penalty factors (λ_1 and λ_2) were determined by *k*-fold cross-validation (Shao, 1993; Hjorth, 1994). More specifically, we used one-year-long blocks of observations to form pairs of evaluation and estimation sets, and then we computed the PRESS (prediction error sum of squares) statistic

$$\sum_{t} \sum_{(t,j)\notin E_t} (y_t^{(j)} - \hat{\alpha}_t^{(j)} - \hat{\beta}_1^{(j)} x_{1t}^{(j)} - \ldots - \hat{\beta}_p^{(j)} x_{pt}^{(j)})^2 \qquad (4.3)$$

where E_t depicts the *t*th evaluation set, and the model parameters were estimated using all observations belonging to the complement of the evaluation set. Finally, λ_1 and λ_2 were selected in such a way that the PRESS value was minimised. In a recently published comparison of *k*-fold and leave-one-out cross-validation (Libiseller and Grimvall, 2003), the former method was found to entail less risk for over-fitting when the residuals were serially correlated.

In Chapter 2, it was noted that the semiparametric model (4.1) for data collected over several seasons can be regarded as a special case of a more general model for trend assessment of multivariate time series. In fact, only the second roughness penalty term in 4.2 must be modified to achieve the circular and gradient smoothings discussed in Chapter 2. Moreover, if we let $s = (t - 1)\, m + j$ and introduce the notation $\alpha(s) = \alpha_t^{(j)}$, that term can be written in the general form

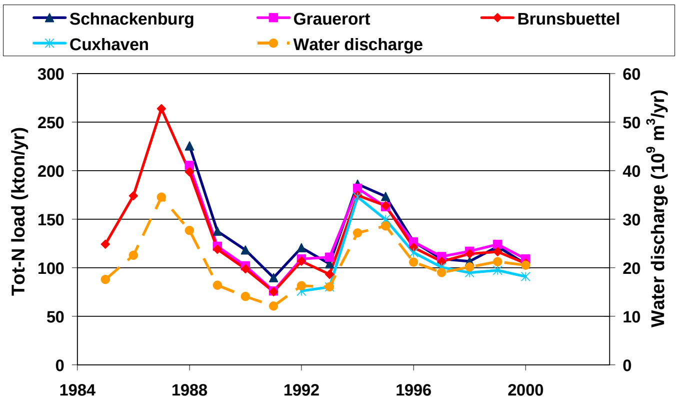
$$\lambda_2 \sum_{s \in \boldsymbol{S}} (\alpha(s_1) - \frac{\alpha(s_2) + \alpha(s_3)}{2})^2 \qquad (4.4)$$

where $\boldsymbol{s} = (s_1, s_2, s_3)$ is a triplet of distinct integer and $\boldsymbol{S}$ is a set of such triplets.

4.2 Normalisation of environmental data using SPR models

Paper IV describes a study of nutrient loads in the Elbe River in Germany, and how such loads can be normalised with respect to water discharge and other factors. Figure 4.1 shows the estimated annual loads of total nitrogen and total phosphorus at one sampling site (Schnackenburg) upstream of Hamburg city, and three sampling sites (Grauerort, Brunsbüttel, and Cuxhaven) downstream of that city, along with annual water discharge values from the sampling site NeuDarchau, which is located downstream of Schnackenburg. A downward tendency can be observed in most of the time series of nutrient loads. However, the interannual variation is large, and the annual loads vary markedly with the annual water discharge values.

a)



b)

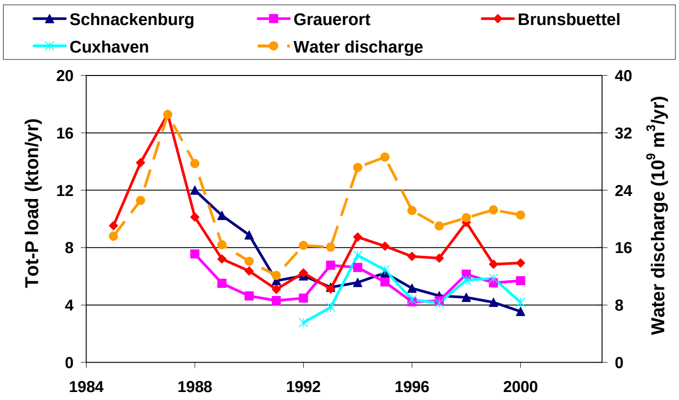


Figure 4.1. Estimated annual loads of nitrogen (a) and phosphorus (b) for different sampling sites shown together with water discharge values from NeuDarchau on the Elbe River.

Figure 4.2 illustrates the same data as in Figure 4.1 after normalisation. The downward trends now emerge much more clearly, and the normalisation was successful in that the interannual variation in the normalised loads is much smaller than in the time series of observed loads. The nitrogen loads were influenced primarily by the amount of water discharge, whereas the phosphorus loads were related not only to the water discharge, but also largely to the load of suspended particulate matter. Figure 4.2 illustrates the results obtained using normalisation models that were found to be optimal (see paper IV).

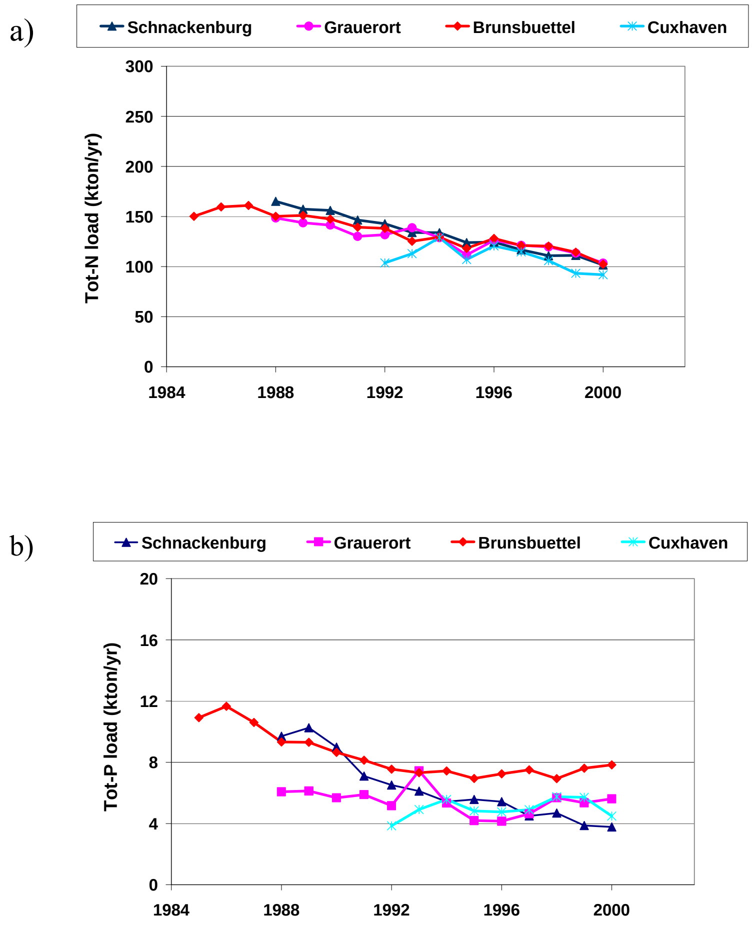


Figure 4.2. Normalised annual loads of total nitrogen at four investigated sampling sites on the Elbe River.

5 Comparison of different regression methods

The regression-based methods that have been developed and utilised in this thesis can be regarded as complements to existing parametric, semiparametric, and nonparametric methods. This raises the question of under what circumstances our techniques will be more suitable than other methods. The following discussion is divided into two parts, which consider annual data and seasonal data, respectively.

5.1 Methods for annual data

One of the main reasons for using MR and other nonparametric approaches is that they do not involve strong assumptions about the relationship that is implicit in standard parametric regression. That property is also seen in the generalised additive models (GAMs), which have the form

$$g\left(E(y \mid x_1, ..., x_p)\right) = \alpha + h_1(x_1) + ... + h_p(x_p) \tag{5.1}$$

where g is what is known as a link function, and the functions h_k are estimated using some kind of scatter plot smoother (Hastie *et al.*, 2001). Some models that are even more general are usually also referred to as GAMs, for example, those of the form

$$g\left(E(y \mid x_1, ..., x_p)\right) = \alpha + h(x_1, x_2) \tag{5.2}$$

which allow interaction effects between two predictors.

Let us now consider a data set from a monitoring programme that studied flounder (*Platichthys flesus*) caught in the North Sea (51° 19′ 59″ N, 2° 10′

0″ E) with regard to the concentration of mercury in muscle tissue. Figure 5.1 illustrates how the mercury level is associated with body length and sampling year. The diagram clearly shows that the mercury levels increased with increasing body length, and they also tended to decrease over time.

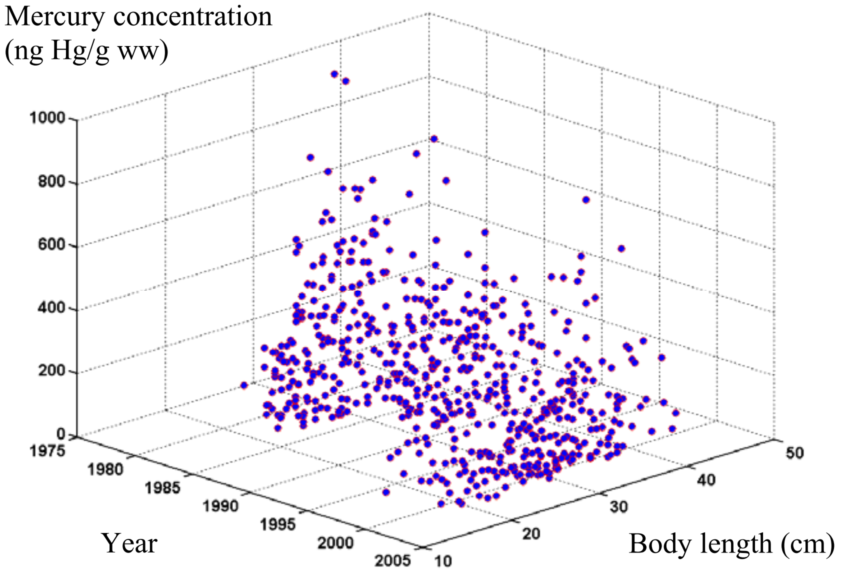
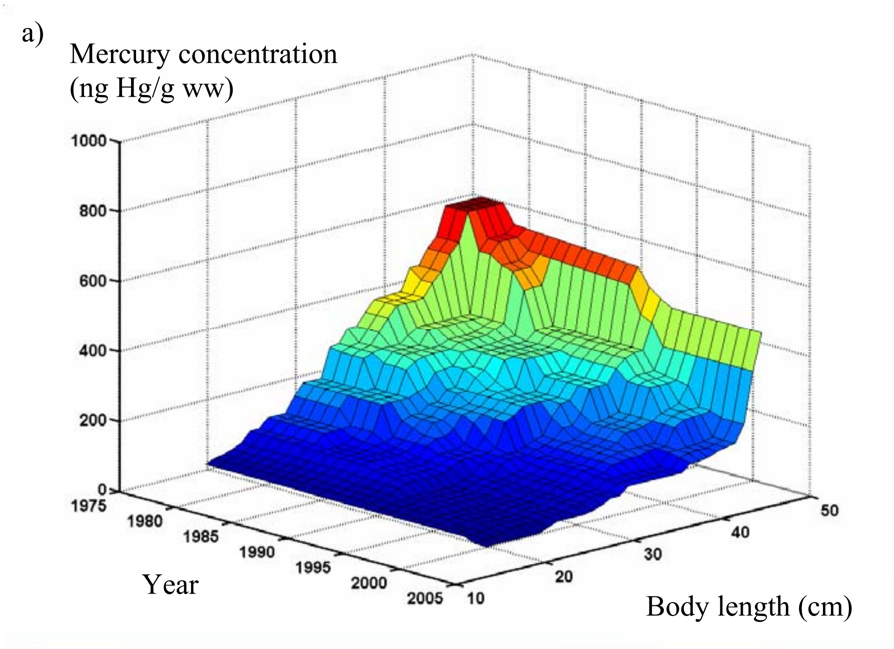


Figure 5.1. Observed concentrations of mercury in muscle tissue from flounder (Platichthys flesus) caught in the North Sea (51^o 19′ 59″ N, 2^o 10′ 0″ E).

Figure 5.2 illustrates the response surfaces obtained when the mercury data were analysed using MR (a), a general additive model (b), and a thin plate spline model allowing for interaction effects of sampling year and body length (c). The additive model produced the least credible response surface, because the analysed data exhibited a strongly nonlinear pattern. The response surfaces generated by monotonic regression and thin plate spline, respectively, appeared to differ mainly in regard to the smoothness of the fitted surface. However, closer examination of the two surfaces revealed that there was also a substantial difference in the estimated mercury trends, particularly for small body sizes. This can be attributed to the fact that thin plate splines tend to smooth out non-additive features in observed data, whereas MR leaves such features unchanged unless they violate the monotonicity.



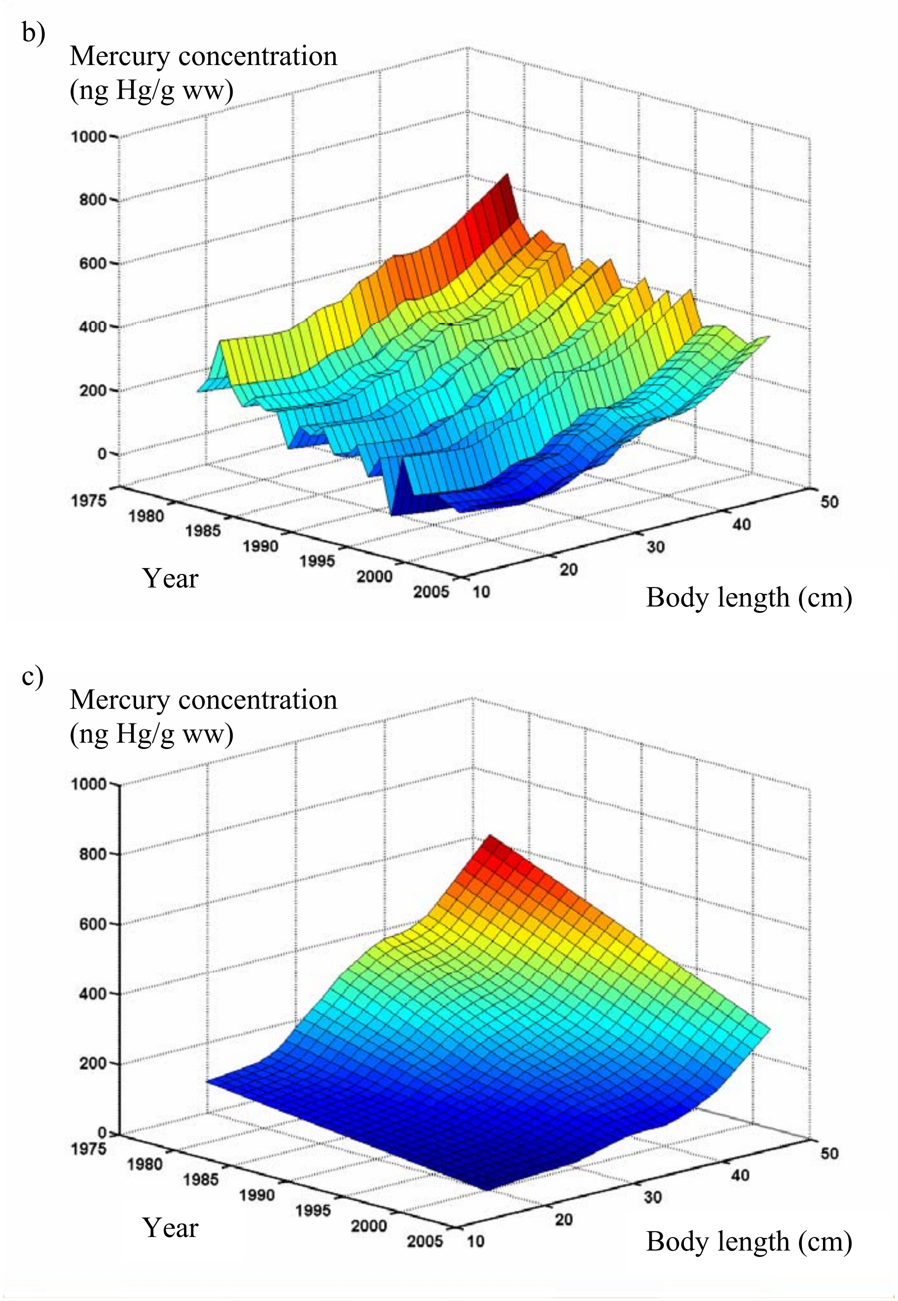
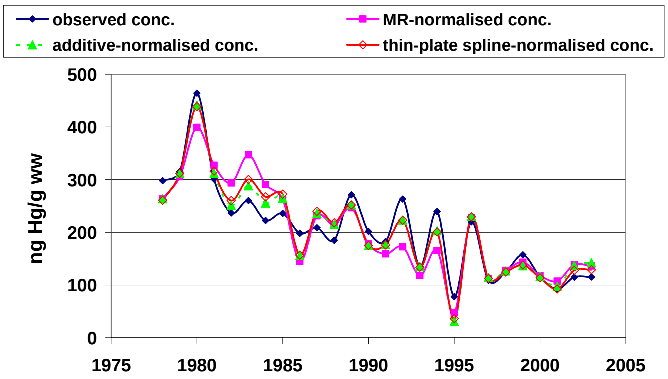


Figure 5.2. Estimated response surfaces for the concentration of mercury in muscle tissue from flounder (Platichthys flesus) caught in the North Sea (51° 19' 59" N, 2° 10' 0" E). Three different estimation were used: monotonic regression (a), general additive model (b), and thin plate spline (c).

When the three models were used to normalise the mercury concentrations with respect to body length, the results obtained were fairly similar (Figure 5.3). This observation could be attributed to the large random variation in collected data that remained even after the normalisation. Further details on this subject are given in paper IV.



5.3. Normalised mercury concentrations in muscle tissue from flounder (Platichthys flesus) caught in the North Sea (51^o 19′ 59″ N, 2^o 10′ 0″ E). The normalisation to a body length of 31.5 cm was based on monotonic regression, a general additive model, and thin plate spline.

5.2 Methods for data collected over several seasons

The SPR method presented in this thesis was developed specifically for the analysis of data collected over several seasons, and as shown in Chapter 3, MR methods can also be applied to such data. Generalised additive models can easily handle additive seasonal effects, although, when the response to covariates varies between seasons, it can be difficult to introduce proper smoothing over seasons for the trend function.

Therefore, let us focus on a comparison of MR and our roughness penalty approach to SPR.

Figure 5.4 shows the monthly loads of nitrogen and water discharge values recorded in the Rhine River at Lobith during the period 1989-2002. A downward trend in the nitrogen loads can be discerned, but the large interannual variation calls for a thorough analysis.

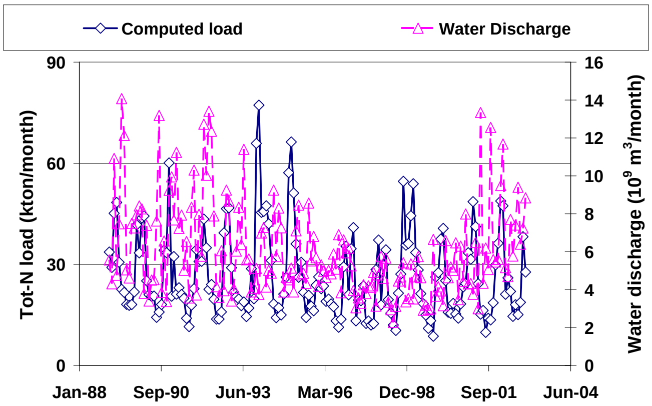


Figure 5.4. Time series plot of monthly riverine loads of total nitrogen in the Rhine River at Lobith.

The scatter plot in Figure 5.5 indicates two things. First, the load seems to increase almost linearly in relation to water discharge. Second, the load for a given level of the water discharge appears to decrease over time. Accordingly, both MR and SPR seem to be applicable. Closer examination of the seasonal pattern indicated that, in the case of MR, it could be assumed that the maximum and minimum occurred in February and August, respectively.

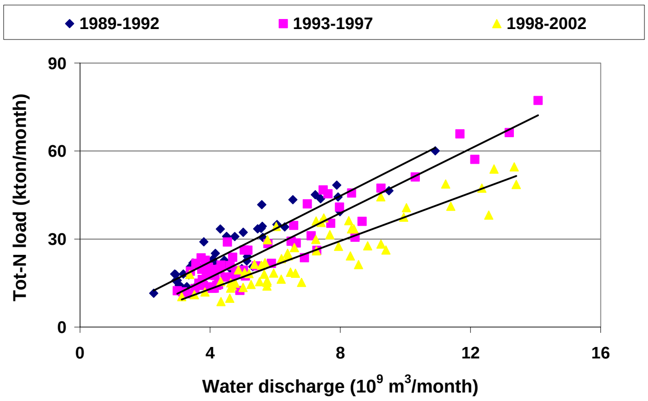


Figure 5.5. Scatter chart of monthly loads of total nitrogen in relation to monthly runoff in the Rhine River at Lobith for different time periods.

The time series plots of flow-normalised nitrogen loads depicted in Figures 5.6 and 5.7 show that both MR and SPR successfully removed most of the interannual variation, and the estimated temporal trends were similar. However, a more comperhensive analysis of the normalised annual loads showed that the normalisation based on MR was slightly more efficient. This was expected, because the scatter plot in Figure 5.5 indicated that the effect of water discharge on nitrogen loads changed over time. Once again it can be emphasised that the occurrence of non-additive patterns in collected data is crucial for the choice of normalisation model.

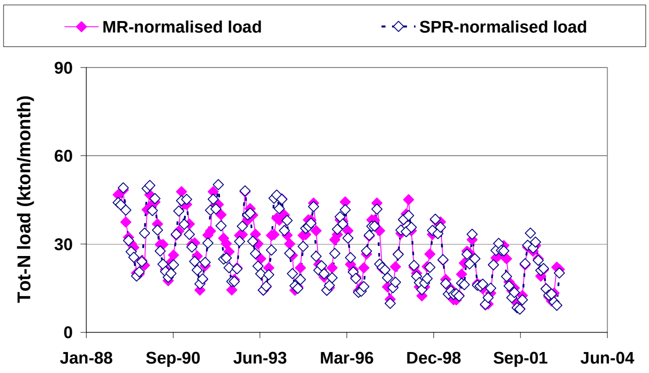


Figure 5.6. Normalised monthly loads of total nitrogen in the Rhine River at Lobith.

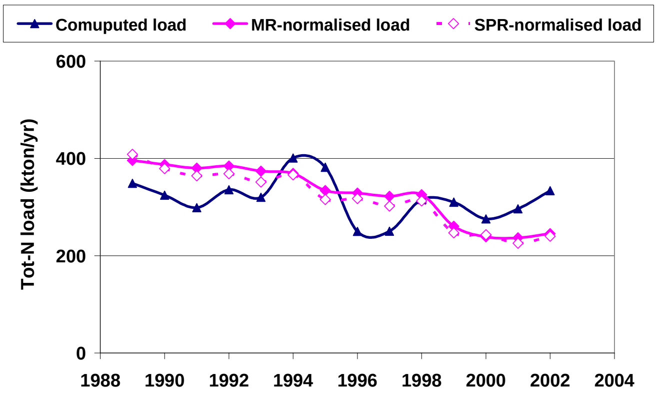


Figure 5.7. Normalised annual loads of total nitrogen in the Rhine River at Lobith.

6 Conclusions and final remarks

The results presented in this thesis show that regression-based normalisation can greatly facilitate the assessment of temporal trends in time series of environmental quality data. To enable flexible modelling of the temporal dynamics of collected data we focused on semiparametric and nonparametric methods. In either case, the trend was modelled in a nonparametric fashion, which rendered the proposed methods particularly suitable for the detection of nonlinear trends.

Normalisation based on MR may appear to be the method of choice when the response variable increases or decreases in relation to time and to one or more variables representing natural variability. There are two main reasons for this. First, it is easy to communicate MR models to non-statisticians, because the basic assumptions usually have simple biogeochemical explanations. Second, our research group at Linköping University has developed computationally efficient methods to undertake MR-based normalisation and trend assessment. The simulation studies reported in this thesis provided additional support for such normalisation. In particular, we found that the performance of the generalised pool adjacent violators (GPAV) algorithm was greatly enhanced when a Hasse diagram was employed to presort the data entered into the calculations. We also found that a Visual Basic implementation of this algorithm could conveniently handle data sets comprising up to 10,000 observations.

Notwithstanding, MR has some weaknesses that call for further investigation. Our simulations demonstrated that the unadjusted mean square residual can strongly underestimate the variance of the error terms, and that the current methods to compensate for the degrees of freedom of the fitted model are inadequate for MR models. Another important aspect is that the techniques for model selection are still in their infancy. We propose that additional work be done to clarify how the degrees of freedom of MR models can be more accurately assessed by taking into account relevant characteristics of a Hasse diagram of the $\boldsymbol{x}$ vectors. Also, it would be very interesting to determine how the present algorithms for monotonic regression can be further developed to facilitate model selection by cross-validation.

The semiparametric methods proposed in this thesis were originally developed to handle data collected over several seasons. We generalised an existing algorithm so that it could accommodate several explanatory variables and proposed a block cross-validation procedure for the model selection. However, it is essential to note that these methods have a much wider field of applications. As outlined in Chapter 2, the model for seasonal data can be regarded as a special case of a general model for joint normalisation of multivariate time series of data.

The formal definition of the normalisation concept and the subsequent discussion of various regression models related in Chapter 2 represent an attempt to provide a unified theory for all regression-based normalisation methods. In particular, it should be pointed out that the results described in this thesis make the concepts of global and local normalisation more precise.

However, further work is needed to fully integrate the statistical normalisation techniques discussed here with models that involve explicit information about the anthropogenic forcing of the system under consideration. Above all, efforts should be made to diminish the gap between statistical and process-based deterministic modelling for trend assessment of environmental quality data.

The case studies of mercury in North Atlantic cod and nutrient loads carried by the Elbe River illustrate the applicability of the proposed normalisation techniques. The Elbe investigation was particularly successful in the sense that normalisation almost completely removed the irregular year-to-year variation and the models obtained by cross-validation were consistent with the knowledge gained from process-based studies of riverine nutrient loads. The results of the mercury study were less clear, probably because such data can be strongly influenced by measurement errors. The example of MR of total nitrogen on time and salinity (Chapter 3) indicates that data on water quality in estuaries and archipelagos are particularly suitable for normalisation, and further work in this context is in progress.

References

Ayer M., Brunk H. D., Ewing G. M., Reid W. T. and Silverman E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, **26**, 641-647.

Barlow R. E., Bartholomew D. J., Bremner J. M. and Brunk H. D. (1972). *Statistical inference under order restrictions*. New York, Wiley.

Best M. J. and Chakravarti N. (1990). Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming*, **47**, 425-439.

Bloomfield P., Royle J. A., Steinberg L. J. and Yang Q. (1996). Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmospheric Environment*, **30**, 3067-3077.

Bril G., Dykstra R., Pillers C. and Robertson T. (1984). Algorithm AS 206, isotonic regression in two independent variables. *Applied Statistics*, **33**, 352-357.

Burdakov O., Grimvall A. and Hussian M. (2004). *A generalised PAV algorithm for monotonic regression in several variables*. In: Antoch, J. (ed.) COMPSTAT, Proceedings of the 16th Symposium in Computational Statistics held in Prague. Heidelberg, New York, Physica-Verlag (Springer), 761-767.

Burdakov O., Sysoev O., Grimvall A. and Hussian M. (2005). *An $O(n^2)$ algorithm for isotonic regression*. In: G. Di Pillo and M. Roma (Eds.) *Large Scale Nonlinear Optimization*. Heidelberg, Springer-Verlag, 25-33.

Clark M. W., Davies F., McConchie M. D. and Birch G. F. (2000). Selective chemical extraction and grainsize normalisation for environmental assessment of anoxic sediments: validation of an integrated procedure. *The Science of the Total Environment*, **258**, 149-170.

Davey B. A. and Priestly H. A. (2002). *Introduction to Lattices and Order.* Cambridge, Cambridge University Press, 2nd edition.

De Simone V., Marino M. and Toraldo G. (2001). In*:* Floudas, C. A. and Pardalos, P. M. (Eds.) *Encyclopedia of optimization*. Dordrecht, Kluwer Academic Publishers.

Dykstra R. and Robertson T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, **10**, 708-716.

Forsman Å. and Grimvall A. (2003). Reduced models for efficient simulation of spatially integrated outputs of one-dimensional substance transport models. *Environmental Modelling and Software*, **18**, 319-327.

Gardner M. W. and Dorling S. R. (2000a). Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment*, **34**, 21-34.

Gardner M. W. and Dorling S. R. (2000b). Meteorologically adjusted trends in UK daily maximum surface ozone concentrations. *Atmospheric Environment,* **34**, 171-176.

Giannitrapani M., Bowman A. W. and Scott E. M. (2005). Additive models for correlated data with applications to air pollution monitoring. *Submitted to Biometrics*.

Green P. J. and Silverman B. W. (1994). *Nonparametric regression and generalised linear models - a roughness penalty approach*. London, Chapman and Hill.

Grimvall A., Wackernagel H. and Lajaunie C. (2001). *Normalisation of environmental quality data.* In: L.M. Hilty and P.W. Gilgen (Eds.) Sustainability in the Information Society. Marburg, Metropolis-Verlag, 581-590.

Hanson D. L., Pledger G. and Wright F. T. (1973). On consistency in monotonic regression. *The Annals of Statistics*, **1**, 401-421.

Härdle W. (1997). *Applied non-parametric regression*. Cambridge, Cambridge University.

Hastie T., Tibshirani R. and Friedman J. (2001). *The elements of statistical learning*. New York, Springer.

Hirsch R. M., Slack J. R. and Smith R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, **18**, 107-121.

Hjorth U. (1994). *Computer intensive statistical methods.* London, Chapman & Hall.

Holland D. M., Principe P. P. and Vorburger L. (1999). Rural ozone: trends and exceedances at CASTNet sites. *Environmental Science & Technology*, **33**, 43-48.

Huang L. S. and Smith R. L. (1999). Meteorologically-dependent trends in urban ozone. *Environmetrics*, **10**, 103-118.

Koelmans A., Gillissen F., Makatita W. and Van Den Berg M. (1997). Organic carbon normalisation of PCB, PAH and pesticide concentrations in suspended solids. *Water Research*, **31**, 461-470.

Libiseller C. and Grimvall A. (2003). Model selection for local and regional meteorological normalisation of background concentrations of tropospheric ozone. *Atmospheric Environment*, **37**, 3923-3931.

Libiseller C., Grimvall A. and Hussian M. (2005). Impact of improved wastewater treatment on the concentration of total nitrogen in the Stockholm archipelago. *Research Report LIU-MAI-R-2005-07*, *Department of Mathematics, Linköping University*, Linköping, Sweden.

Libiseller C., Grimvall A. and Hallberg L. (2005). Meteorological normalisation of time series of wet deposition. *Manuscript.*

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

Mallows C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362-372.

Mayer M. and Woodroofe M. (2000). On the degrees of freedom in sharp-restricted regression. *The Annals of Statistics*, **28**, 1083-1104.

Mukarjee H. (1988). Monotone nonparametric regression. *The Annals of Statistics*, **16**, 741-750.

Mukarjee H. and Stern H. (1994). Feasible nonparametric estimation of multiargument monotone functions. *Journal of the American Statistical Association*, **425**, 77-80.

Pardalos P. M. and Xue G. (1999). Algorithms for a class of isotonic regression problems. *Algorithmica*, **23**, 211-222.

Robertson T. and Waltman P. (1968). On estimating monotone parameters. *The Annals of Mathematical Statistics*, **39,** 1030-1039.

Salanti G. and Ulm K. (2001). *The multidimensional isotonic regression*. Proceedings of the International Society of Clinical Biostatistics, 19-23 Aug 2001, Stockholm, p. 162.

Schell M. J. and Singh B. (1997). The reduced monotonic regression method. *Journal of the American Statistical Association*, **92**, 128-135.

Schimek M. G. (2001). *Smoothing and Regression: Approaches, Computation, and Application*. New York, Wiley-Interscience.

Shao J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486-494.

Shively T. S. and Sager T. W. (1999). Semiparametric regression approach to adjusting for meteorological variables in air pollution trends. *Environmental Science & Technology*, **33**, 3873-3880.

Stålnacke P. and Grimvall A. (2001). Semiparametric approaches to flow-normalisation and source apportionment of substance transport in rivers. *Environmetrics,* **12**, 233-250.

Stålnacke P., Grimvall A., Sundblad K. and Wilander A. (1999). Trends in nitrogen transport in Swedish rivers. *Environmental Monitoring and Assessment,* **59**, 47-72.

Strand M. (2003). Comparisons of methods for monotone nonparametric multiple regression. *Communications in Statistics, Simulation and Computations*, **32**, 165-178.

Thompson M. L., Reynolds J., Cok L. H., Guttorp P. and Sampson P. D. (2001). A review of statistical methods for the meteorological adjustment of ozone. *Atmospheric Environment,* **35**, 617-630.

Uhlig S. and Kuhbier P. (2001). *Trend methods for the assessment of the effectiveness of reduction measures in the water system*. Umweltforchungsplan (UFOPLAN), Nr. 298 22 244.
Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit, Berlin.

Wahlin K., Shahsavani D., Grimvall A., Wade A. and Butterfield D. (2004). *Reduced models of the retention of nitrogen in catchments.* In: C. Pahl-Wostl, S. Schmidt, A.E. Rizzoli, and A. J. Jakeman (Eds.) Complexity and Integrated Resources Management, Transactions of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society, iEMSs: Manno, Switzerland.