Roadmap for Trend Detection and Assessment of Data Quality

Karl Wahlin

Linköping University
FACULTY OF ARTS AND SCIENCES

Linköping Studies in Statistics No. 10
Linköping Studies in Arts and Science No. 454
Linköping University, Department of Computer and Information Science
Linköping 2008

Linköping Studies in Statistics No. 10 • Doctoral Thesis
Linköping Studies in Arts and Science No. 454 • Doctoral Thesis

At the Faculty of Arts and Science at Linköping University, 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 Science. This thesis comes from the Division of Statistics at the Department of Computer and Information Science.

Distributed by:
Department of Computer and Information Science
Division of Statistics
Linköping University
SE-581 83 Linköping, Sweden

Karl Wahlin
Roadmap for Trend Detection and Assessment of Data Quality

Edition 1:1
ISBN: 978-91-7393-792-4
Linköping Studies in Statistics, ISSN 1651-1700
Linköping Studies in Arts and Science, ISSN 0282-9800

Printed by LiU-Tryck, Linköping 2008

Abstract

Regular measurements of the state of the environment constitute a cornerstone of environmental management. Without the support of long time series of reliable data, we would know much less about changes that occur in the environment and their causes. The present research aimed to explore how improved techniques for data analysis can help reveal flawed data and extract more information from environmental monitoring programmes. Based on our results, we propose that the organization of such monitoring should be transformed from a system for measuring and collecting data to an information system where resources have been reallocated to data analysis. More specifically, this thesis reports improved methods for joint analysis of trends in multiple time series and detection of artificial level shifts in the presence of smooth trends. Furthermore, special consideration is given to methods that automatically detect and adapt to the interdependence of the collected data. The current work resulted in a road map describing the process of proceeding from a set of observed concentrations to arrive at conclusions about the quality of the data and existence of trends therein. Improvements in existing software accompanied the development of new statistical procedures.

Papers included in this thesis

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

I Grimvall A., Wahlin K., Hussian M. and von Brömssen C. (2008). Semiparametric smoothers for trend assessment of multiple time series of environmental quality data. Submitted to *Environmetrics*.

II Wahlin K. and Grimvall A. (2008). Uncertainty in water quality data and its implications for trend detection: lessons from Swedish environmental data. *Environmental Science & Policy* **11**, 115-124.

III Wahlin K. and Grimvall A. (2008). Roadmap for assessing regional trends in groundwater quality. Submitted to *Environmental Monitoring and Assessment*.

IV Wahlin K., Grimvall A. and Sirisack S. (2008). Estimating artificial level shifts in the presence of smooth trends. Submitted to *Environmental Monitoring and Assessment*.

V Wahlin K., Shahsavani D., Grimvall A., Wade A.J., Butterfield D. and Jarvie H.P. (2004). Reduced models of the retention of nitrogen in catchments. In *Proceedings of the International Environmental Modelling and Software Society Conference* (iEMSs), Osnabrück, Germany, 14–17 June, 2004.

My contribution to the papers

I had the main responsibility for Papers II-V. In Paper I, I was responsible for the development of the new resampling procedure.

Acknowledgements

First and foremost, I want to express my gratitude to my supervisor, Professor Anders Grimvall, for his invaluable support during my work. I also want to thank Dr. Anders Nordgaard and Dr. Claudia von Brömssen for constructive comments on drafts of this thesis. I am obliged to Ms. Patricia Ödman for kind help with the language revision. A special thanks to all former and present colleagues at the Division of Statistics for assistance and encouragement. Furthermore, I am grateful for funding from the Swedish Environmental Protection Agency and the Geological Survey of Sweden.

Mats, Gun, Lotta, and Hala: you are truly special.

Karl Wahlin
Linköping, September 2008

Contents

1 Introduction

1.1 Motivation and scientific context

Regular measurements of the state of the environment make up the foundation of evidence-based environmental management. Without the existence of long time series of reliable data collected in local, regional, or national monitoring programmes, we would know much less about the changes that occur in the environment and their causes. Also, it would be considerably more difficult to assess the efficacy of measures taken to combat the deterioration of ecosystems. However, the fact that monitoring is an indispensable part of environmental management does not necessarily imply that it is coordinated in an optimal manner. In almost all organizations, there is a substantial risk that views and methodologies are not being updated to meet new demands and take advantage of novel technologies.

At the time many of the existing monitoring programmes were devised, it was not unusual to address important questions by performing simple surveys and very basic data analyses. Hot spots and particularly vulnerable environments were successfully delineated, and the impact of removing major point emissions was often readily documented in a convincing fashion. The environmental issues that are now receiving the most attention are more complex and involve evaluation of much larger datasets. Regional or global change is usually in focus, and understanding long-term alterations in the state of the environment represents a priority. In addition, the human impact that a monitoring system is intended to detect is often relatively small compared to both the natural interannual variation at the monitored sites and the random measurement errors that influence the individual observations. The objective of the research underlying this thesis was to explore how improved techniques for data analysis can help reveal flawed data and how more information can be extracted from environmental monitoring programmes. The examples discussed here originate from water quality monitoring, but the major conclusions are valid for a much larger class of monitoring programmes.

On a more general level, the current results raise questions about the need for new paradigms in environmental monitoring. Over the past decade, it has gradually become more apparent that science and professional work increasingly entail collection, organization, transformation, and presentation of information. Together, powerful computers with almost unlimited storage capacity and the Internet and its efficient search engines constitute much more than a technological revolution. Dramatic changes are also occurring in the organization and performance of science and professional work. In a recently published issue of the journal *Nature*, it was stressed that “important discoveries are made by scientists and teams who combine different skill sets—not just biologists, physicists and chemists, but also computer scientists, statisticians and data-visualization experts” (Szalay & Gray, 2006, p. 413). Moreover, another author in that issue claimed that “It will be a very different way of thinking, sifting through the data to find patterns” (Butler, 2006, p. 403). It was even stated that “applied computer science is now playing the role which mathematics did from the seventeenth through the twentieth centuries: providing an orderly, formal framework and exploratory apparatus for other sciences” (Foster, 2006, p. 19). The discussion here deals with how the mentioned development can and should influence the extraction of information from environmental monitoring programmes. In particular, the focus is on the role of statistical methods that are sufficiently simple so as to be easy to comprehend, yet sufficiently advanced to capture the key features of the collected data.

From a statistical point of view, our work has its roots in a visionary article by Tukey (1962). Long before the width of the present revolution of computer science and technology could be anticipated, Tukey asserted the following:

- Data analysis is a larger and more varied field than statistical inference in a given probability model.
- There is a need for free use of ad hoc and informal procedures in seeking indications—*"listening to the data."*
- Model building is an iterative procedure.

Tukey also emphasized that the simple graph has brought more information to the mind of the data analyst than has any other computational device.

Nowadays, iterative methods for data analysis are widely used, because they enable data-driven procedures that automatically adapt to the structure of the collected data. We believe that assessment of data quality should also be treated as an iterative process. It is not unusual that data that are considered to be correct at the time of the sampling are deemed erratic after more data have been collected and new analytical procedures have been implemented.

1.2 Study objectives and methods

As indicated in the above-mentioned motivation for the present research, the work had two major objectives:

(i) to facilitate the detection of trends and assessment of data quality in environmental monitoring;

(ii) to demonstrate the need for reallocating resources from data collection to data analysis.

Detection of trends in the state of the environment requires adequate methods to handle outliers, artificial level shifts, temporal and spatial correlation in the collected data, missing values, and observations below the limit of detection or quantification. We investigated how multiple time series of data can be analysed in the presence of such peculiarities. Special attention was paid to non- and semiparametric approaches for joint evaluation of multiple time series. Joint evaluation of data representing many locations and time points is crucial when examining regional and large-scale trends. Non- and semiparametric models are well suited for data-driven and interactive approaches that help the user "listen to the data."

At the onset of our studies, we focused on extraction of smooth trends that could be interpreted as human impact on the environment. However, as our work progressed, we observed that artificial level shifts not related to actual changes in the environment were a major source of long-term temporal variation in the analysed data. This triggered development of statistical methods and software for joint assessment of smooth trends and artificial level shifts in vector time series.

In any discussion of environmental monitoring, it is also necessary to address the balance between process-based modelling and the collection and analysis of observational data. Hence, this thesis provides a brief account of how the two activities can support each other and how statistical analysis of model inputs and outputs can provide valuable information for environmental management.

To be more precise, the description covers the development and application of three main methods:

(i) *Extended Mann-Kendall tests for monotonic trends*

Mann-Kendall tests for monotonic trends have long been among the major instruments used to detect temporal trends in environmental data (Hirsch & Slack, 1984). We examine how such tests can be extended to simultaneously handle serial correlation, covariates, and censored data. In addition, to support such analyses, we modify existing software.

(ii) *Semiparametric regression for the detection of smooth trends and artificial level shifts in vector time series*

In this work, we generalize models that had previously been used to assess smooth trends in multiple time series of data (Hussian *et al*., 2004; Stålnacke & Grimvall, 2001). In particular, we examine how artificial level shifts can be estimated in the presence of smooth trends and how an existing resampling technique for uncertainty assessments can be modified to accommodate data that are correlated in time or across coordinates in the analysed vector time series.

(iii) *Ensemble runs for extracting the essence of complex process-based models*

Management objectives are typically expressed as long-term spatial averages, whereas process-based models operate on spatial and temporal scales that facilitate mechanistic interpretation of equations and model parameters. The text here briefly describes how multiple runs of a process-based model of nitrogen flows in catchments can be used to extract important features of temporally aggregated model outputs and how such model runs can support assessment of trends in observational data.

An important aim of this thesis is to integrate the first two methods into a roadmap for trend detection and assessment of data quality. In addition, we develop two software packages (Multitest and Multitrend) to support this roadmap. Multitest aims to achieve a preliminary check of trends in the observed time series, whereas Multitrend is devoted to revealing synchronous increases and decreases and to separating smooth trends from abrupt level shifts.

1.3 Outline of the thesis

This summary is based on five papers (designated I–V), which are appended at the end of the dissertation.

Paper I is dedicated to semiparametric smoothing and the major ideas behind Multitrend. In the next study (Paper II, published in *Environmental Science & Policy* 11, pp. 115–124), Multitrend is used to investigate water quality trends in Swedish water courses and to scrutinize the quality of collected data. Thereafter (Paper III), Multitest and Multitrend are used to investigate Swedish groundwater data and to demonstrate how trends can be extracted from large datasets involving temporally and spatially correlated observations. During the work underlying Papers II and III, serious data quality problems were revealed. In particular, we found that artificial level shifts had a marked impact on long-term trends. Therefore, we conducted an additional study (Paper IV) in which Multitrend was extended to enable detection and estimation of abrupt changes in vector time series. Paper V (published in the *Proceedings of the International Environmental Modelling and Software Society Conference*, Osnabrück, Germany, 14–17 June, 2004) provides an example of the need for better integration of process-based modelling and statistical data analysis.

The final chapter of this thesis summarizes the conclusions drawn from our research and addresses the need for a paradigm shift in environmental monitoring.

2 Testing for trends in multiple time series

Assessment of temporal trends in data representing a network of stations requires statistical methods that can accommodate multiple, statistically dependent time series. In addition, it is highly desirable that the selected methods are easy to comprehend, are robust to outliers, and can accommodate missing values. This makes the Mann-Kendall (MK) tests advocated by Hirsch and Slack (1984) an attractive choice. Such nonparametric techniques also have the advantage of being less demanding for the user compared to parametric methods that require detailed modelling of probability distributions and dependencies.

The aim of statistical trend testing is to distinguish between deterministic and stochastic changes over time, and thus the character of the stochastic components is always a key issue. In particular, it is necessary to take into account the presence of serial correlation. Hirsch and Slack (1984) showed how dependence across seasons can be handled. However, there is no ideal method to adjust for serial dependence over time intervals longer than a year, and existing methods are also restricted to univariate time series (Hamed & Rao, 1998; Yue & Wang, 2004). Here, we examined a simple procedure based on reorganization of the given data into a larger number of time series with longer time steps. In addition, we clarified that censored data can easily be handled in all kinds of MK tests, including partial MK tests involving covariates (Libiseller & Grimvall, 2002).

Inasmuch as the collected data usually represent things like several seasons, stations, regions, or sampling methods, it is often of interest to test for trends in several subgroups of data. This calls for software in which multiple tests are automatically undertaken. The software Multitest, which we modified in the present work, provides such features for both ordinary and partial MK tests.

2.1 The ordinary univariate MK-test

The classical MK method is a nonparametric test for monotone trend in a single time series $y_1, \ldots, y_n$. It is based on pairwise comparisons of observations, and the test statistic can be written

$$T = \sum_{i<j} \operatorname{sgn}(y_j - y_i)$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1, \text{ if } x > 0 \\ 0, \text{ if } x = 0 \\ -1, \text{ if } x < 0 \end{cases}$$

Under the assumption that there is no trend in the data (i.e., that all permutations of the observed values are equally likely), the distribution of T can be approximated by a normal distribution with mean zero and variance $n(n-1)(2n+5)/18$. When missing values or ties occur, the variance is smaller, and the formulae presented by Hirsch and Slack (1984) automatically adjust for such events.

2.2 Partial MK-tests

The ordinary MK test can be generalized to enable testing for a trend while simultaneously adjusting for a trend in a covariate. Let T and S denote the test statistics for trend in the response and covariate, respectively. Then, the test statistic of the partial MK test can be written

$$U = \frac{T - \hat{\rho}_{T,S} S}{\sqrt{\hat{V}(T)(1 - \hat{\rho}_{T,S}^2)}}$$

where $\hat{V}(T)$ denotes the estimated variance of T, and $\hat{\rho}_{T,S}$ represents the estimated correlation of T and S (El-Shaarawi & Niculescu, 1992; Libiseller & Grimvall, 2002). Under the null hypothesis that there is no trend that can be attributed to factors other than a trend in the covariate, the test statistic is approximately normal with mean zero and variance one.

2.3 Multivariate MK-tests

When investigating a geographical region, different sampling sites often exhibit similar, albeit not identical, trends. This calls for tests in which the evidence of trends is pooled according to user-defined grouping criteria. Hirsch and Slack (1984) considered test statistics of the form

$$T = T_1 + ... + T_m$$

where T_1, …, T_m denote the ordinary MK statistics for the individual time series. If the collected data are organized in a matrix so that each row represents a sampling occasion, and each column signifies a sampling site, a season, or some other group of observations, the null hypothesis implies that all row permutations are equally likely. The columns are usually statistically dependent, and Hirsch and Slack (1984) showed how the variance of T under the null hypothesis can be estimated in the presence of such dependencies.

If the collected data can be grouped according to p factors (e.g., sampling sites, seasons, or regions), there are a total of 2^p-1 sum tests in which univariate test statistics are summed over all levels of a subset of factors. However, many of these tests can be redundant. For instance, summation over regions for a given station will create redundant tests, because each station belongs to a single region. In our Multitest software, tests are automatically performed for all non-redundant sum tests.

2.4 Multivariate partial MK tests

Multivariate partial MK tests have been developed for assessing the presence of joint trends in several groups of data. Letting T and S denote test statistics from sum tests for trends in the response and covariate, respectively, the analytical expression of the test statistic is identical to that of the univariate test. Further details about partial MK tests have been reported by Libiseller and Grimvall (2002).

2.5 MK-tests and censored data

Observations below the limit of detection or quantification carry information that can and should be exploited in trend tests (Helsel, 2005). In the present studies, we simply regarded all observations as intervals. If the measured response could be quantified, both the lower and the upper end of the interval were set to this observed response. If the response was below the limit of quantification, the interval ranged from zero to that limit. Furthermore, we introduced a generalized sign function

$$\operatorname{sgn}(a_i, b_i, a_j, b_j) = \begin{cases} 1, \text{ if } b_i < a_j \\ -1, \text{ if } b_j < a_i \\ 0, \text{ otherwise} \end{cases}$$

where $[a_i, b_i]$ and $[a_j, b_j]$ denote the intervals assigned to a pair of observations. Using this generalized sign function, the test statistics in ordinary and partial MK tests could be computed as usual. Analogously, we defined the Theil slope of the trend as the median of all ratios $\frac{b_j - a_i}{j - i}$ and $\frac{a_j - b_i}{j - i}$ for $i < j$.

2.6 Adjustment for serial correlation

The methods currently used to accommodate serial correlation in MK tests are based on estimation of autocorrelations and adjustment of the variance of the test statistics (Hamed & Rao, 1998; Yue & Wang, 2004). Such techniques perform satisfactorily when tested on artificial data generated by an autoregressive model, but they also have substantial weaknesses. In particular, it should be noted that the estimates of autocorrelations and correction factors are strongly influenced by outliers and trends in the measured data, and removal of outliers and trends prior to the estimation of autocorrelations is a difficult task.

We created a method that is a simple extension of the procedure proposed by Hirsch and Slack (1984) to handle correlations across seasons. First we considered a dataset comprising observations $y_1, \ldots, y_{2n}$ made on $2n$ consecutive years. Then we regarded it as observations from n two-year periods, and reorganized the data in the following matrix:

$$\begin{array}{ccc} \textit{Two-year period} & \textit{First response} & \textit{Second response} \\ 1 & y_1 & y_2 \\ 2 & y_3 & y_4 \\ . & . & . \\ . & . & . \\ n & y_{2n-1} & y_{2n} \end{array}$$

Finally, we substituted the ordinary MK test for a sum test based on the two new columns that were formed. Analogously, one can reorganize m columns of responses into $2m$ columns of responses with doubled time steps. For example, monthly data given in twelve columns with time step one year can be reorganized into 24 columns with time step two years. Of course, it is also possible to reorganize data into periods of three years or more.

Figures 2.1 and 2.2 show how our procedure performed when the original data were observations of a first order autoregressive (AR) model with normally distributed error terms. As expected, there was a considerable loss of power when the new matrices contained a small number of rows (less than ten). However, the two diagrams also show that our procedure can increase the robustness to serial dependence without any serious loss of power. In addition, it is easy to comprehend, and it can be applied to all types of MK and partial MK tests, which makes it an attractive technique for trend assessments.

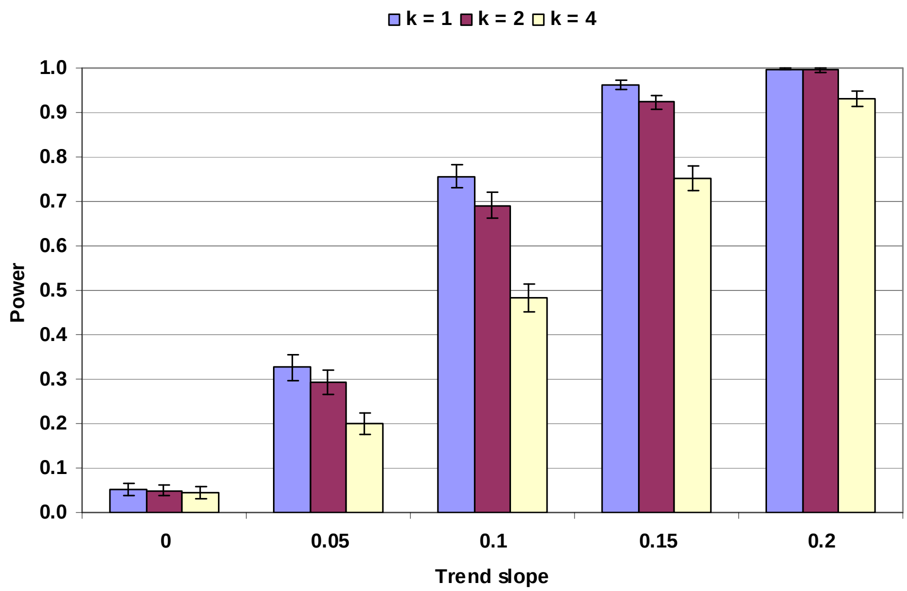


Figure 2.1 Power functions of MK tests when the original 20-year data series was split into *k* series with a time step of *n/k*. Raw data comprised independent normal random variables with variance one and linear slope from 0 to 0.2. The nominal significance level was 5% (one-sided).

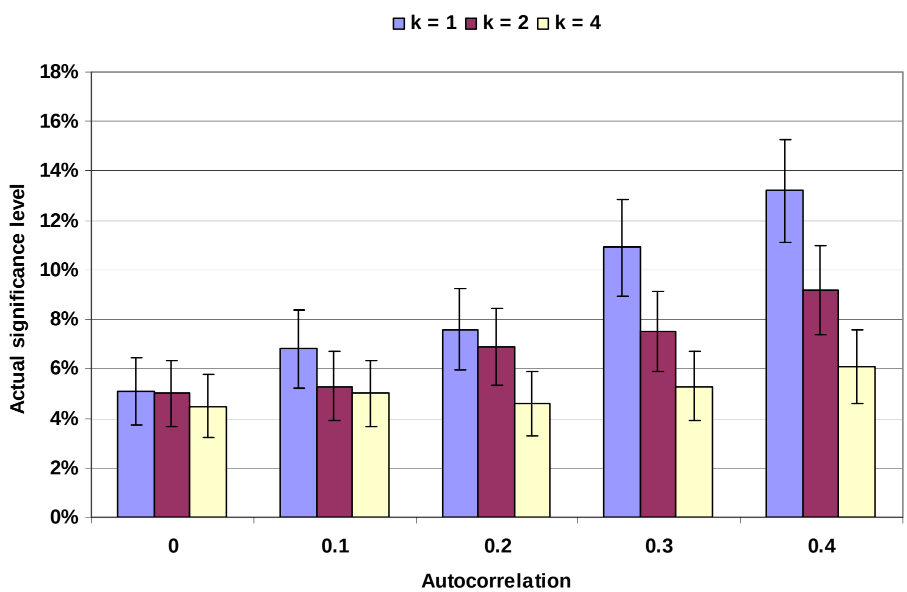


Figure 2.2 Actual significance levels of MK tests based on original and reorganized data when the original series were generated according to AR(1) processes with ρ = 0, 0.1, 0.2, 0.3, and 0.4. The parameter *k* refers to the time step in the reorganized data series, and the nominal significance level was 5% (one-sided).

3 Trend assessment using response surface methodologies

Significance tests for trends provide information about the presence or absence of systematic changes over time. A more detailed trend assessment also requires information about both the shape of the observed trend curves and whether the trends in different groups of data are synchronous. Here, we show how such information can be obtained by fitting response surfaces to vector time series in which the coordinates are ordered according to some user-defined criterion. Figure 3.1 illustrates a trend surface fitted to phosphorus concentrations observed at the mouths of a set of rivers that were ordered with respect to their average phosphorus level.

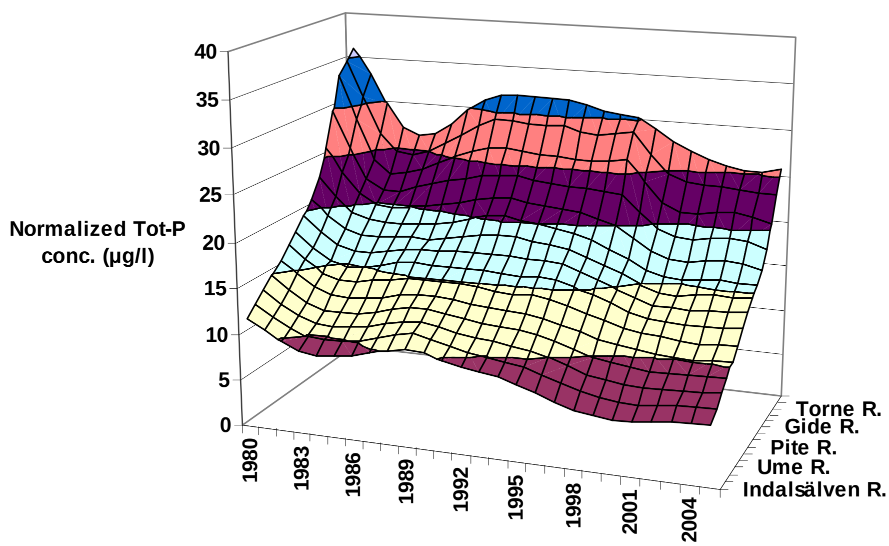


Figure 3.1 Trend surface fitted to normalized annual summaries of total phosphorus (*Tot-P*) concentrations in fifteen Swedish rivers flowing into the Bothnian Bay and Bothnian Sea. The normalization was based on water discharge and the amount of particulate matter, measured as the difference in absorbance between unfiltered and filtered samples. All investigated rivers along with sampling sites are listed in Paper II.

3.1 A basic semiparametric regression model

The origin of our trend surface methodology was a semiparametric regression model developed by Stålnacke and Grimvall (2001) and Hussian and co-workers (2004). This model assumes that all observations can be sorted into a matrix, where the columns correspond to sampling years, and the rows represent sampling sites, seasons, or other groupings of the data. Furthermore, it is assumed that the observed response can be decomposed into the following three components:

(i) a deterministic response surface indicating the human impact on the environment;

(ii) a regression expression describing the impact of covariates representing meteorological variability or other measured natural fluctuations;

(iii) random fluctuations caused by unobservable factors.

Let us first assume that we have exactly one observation for each combination of year and group. Then the observed state of the environment at n equidistant time points will define an m-dimensional vector time series

$$\boldsymbol{y}_t = (y_t^{(1)}, ..., y_t^{(m)})^T, t = 1,...,n$$

and the observations of p covariates representing natural fluctuations of the investigated system can be summarized in a matrix

$$\boldsymbol{x}_t = \begin{pmatrix} x_{1t}^{(1)} & . & . & . & x_{pt}^{(1)} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ x_{1t}^{(m)} & . & . & . & x_{pt}^{(m)} \end{pmatrix}, t = 1, ..., n$$

Furthermore, our response surface model can be expressed as an equation system

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

where the sequence of vectors $\boldsymbol{\alpha}_t = (\alpha_t^{(1)}, ..., \alpha_t^{(m)})^T, t = 1, ..., n$ represents a deterministic temporal trend,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1^{(1)} & . & . & . & \beta_p^{(1)} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ \beta_1^{(m)} & & & & \beta_p^{(m)} \end{pmatrix}$$

is a matrix of time-independent regression coefficients, and the error terms $\varepsilon_t^{(j)}$, j = 1, …, m, t = 1, …, n are identically distributed with mean zero.

The model parameters were estimated by using a penalized least squares technique. For given smoothing factors λ_1 and λ_2, and measures of roughness $L_1(\boldsymbol{\alpha})$ and $L_2(\boldsymbol{\alpha})$ of the intercepts, the parameters were estimated by minimizing

$$S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) + \lambda_1 L_1(\boldsymbol{\alpha}) + \lambda_2 L_2(\boldsymbol{\alpha})$$

where

$$S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{t=1}^{n} \sum_{j=1}^{m} \left(y_t^{(j)} - \alpha_t^{(j)} - \sum_{i=1}^{p} \beta_i^{(j)} \left(x_{it}^{(j)} - \bar{x}_{i.}^{(j)} \right) \right)^2$$

represents the residual sum of squares. Normally we used the expression

$$L_1(\boldsymbol{\alpha}) = \sum_{t=2}^{n-1} \sum_{j=1}^{m} \left(\alpha_t^{(j)} - \frac{\alpha_{t-1}^{(j)} + \alpha_{t+1}^{(j)}}{2} \right)^2$$

to impose smoothing over time whereas $L_2(\boldsymbol{\alpha})$ had different forms for different types of data. When the coordinates of the vector time series were ordered along some gradient, we computed

$$L_2(\boldsymbol{\alpha}) = \sum_{t=1}^{n} \sum_{j=2}^{m-1} \left(\alpha_t^{(j)} - \frac{\alpha_t^{(j-1)} + \alpha_t^{(j+1)}}{2} \right)^2$$

The roughness measure

$$L_2(\boldsymbol{\alpha}) = \sum_{t=1}^{n} \sum_{j=1}^{m} \left(\alpha_t^{(j)} - \frac{\alpha_t^{(j-1)} + \alpha_t^{(j+1)}}{2} \right)^2$$

where

$$\alpha_t^{(m+1)} = \alpha_t^{(1)}$$

and

$$\alpha_t^{(0)} = \alpha_t^{(m)}$$

was used for circular smoothing. Data representing different seasons were smoothed sequentially by setting

$$L_2(\boldsymbol{\alpha}) = \sum_{s=2}^{mn-1} \left(\alpha_s - \frac{\alpha_{s-1} + \alpha_{s+1}}{2} \right)^2$$

where

$$\alpha_s = \alpha_t^{(j)}$$

and

$$s = t(m-1) + j$$

defines the sequential order of the observations.

The gradient, circular and sequential smoothing are special cases of more general smoothing patterns introduced in Paper I. Finally, it can be noted that a small change of the notation in $S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ is sufficient to handle the more general case when the number of observations varies with sampling year or coordinate of the analysed vector time series (Hussian *et al*., 2004).

The methods described above were implemented in our software Multitrend (LiU, 2008). An advantage of our methods over other smoothing methods (mainly Gaussian smoothers and thin plate splines, see Hastie *et al*., 2001; Härdle, 1997) is that the smoothing pattern can be tailored to take into account almost any relationship between the vector components. The numerical algorithms in Multitrend were based on a back-fitting technique suggested by Stålnacke and Grimvall (2001).

3.2 A new resampling technique

Resampling techniques are widely used to estimate the precision of sample statistics. The basic idea behind such techniques is that the distribution of estimators or predictors can be explored by performing simulations or theoretical calculations in which an unknown cumulative distribution function (c.d.f.) is substituted for its empirical c.d.f. The term *bootstrap* is frequently used when new datasets are created by drawing with replacement from a given set of statistically independent observations or model components.

We undertook residual resampling (Mammen, 2000) to assess the precision of parameter estimates in semiparametric regression models. To be more precise, we first computed model residuals

$$e_t^{(j)} = y_t^{(j)} - \hat{\alpha}_t^{(j)} - \sum_{i=1}^{p} \hat{\beta}_i^{(j)} (x_{it}^{(j)} - \overline{x}_{i.}^{(j)}) = y_t^{(j)} - \hat{y}_t^{(j)}, \quad t = 1, ..., n \quad j = 1, ..., m$$

and then assigned new response values to the given predictors by setting

$$y_t^{(j)*} = y_t^{(j)} - e_t^{(j)} + e_t^{(j)*}$$

where $e_t^{(j)*}, t = 1, ..., n,\ j = 1, ..., m$ denotes resampled residuals.

In ordinary residual resampling, new residuals are drawn randomly from the original model residuals, and, consequently, this procedure generates statistically independent error terms. *Block resampling* has been proposed as a means of preserving short-term correlations in time series data (Lahiri, 1999). Furthermore, it has been shown how block resampling procedures can be refined by concatenating with higher likelihood those blocks that match at their ends (Carlstein *et al.*, 1998; Srinivas & Srinivasan, 2005). However, it is unclear how two-dimensional blocks should be selected and matched to achieve optimal results. Therefore, we developed a new form of constrained resampling, in which an ordinary bootstrap sample of the original residuals was modified to restore important dependencies over time and across series.

Constrained residual resampling implied that the resampling favoured those combinations of the original residuals that had desirable auto- and cross-correlations. In the case of gradient smoothing, we first computed all the model residuals and their total variation

$$R_{tot}(\boldsymbol{e}) = \sum_{t=1}^{n}\sum_{j=1}^{m}\left(e_t^{(j)}\right)^2$$

Thereafter, we determined

$$R_1(\boldsymbol{e}) = \sum_{t=2}^{n-1}\sum_{j=1}^{m}\left(e_t^{(j)} - \frac{e_{t-1}^{(j)} + e_{t+1}^{(j)}}{2}\right)^2$$

and

$$R_2(\boldsymbol{e}) = \sum_{t=1}^{n}\sum_{j=2}^{m-1}\left(e_t^{(j)} - \frac{e_t^{(j-1)} + e_t^{(j+1)}}{2}\right)^2$$

where the first sum was used as a measure of the short-term temporal variation of the residuals, and the second one was introduced to capture variation across series. Finally, we formed the ratios

$$T_1(\boldsymbol{e}) = \frac{R_1(\boldsymbol{e})}{R_{tot}(\boldsymbol{e})}$$

and

$$T_2(\boldsymbol{e}) = \frac{R_2(\boldsymbol{e})}{R_{tot}(\boldsymbol{e})}$$

that were subsequently used as targets for a step-by-step modification of ordinary bootstrap residuals. Targets for other smoothing patterns are formed analogously (Paper I).

The modification of bootstrap residuals $\boldsymbol{e}^*$ was based on an iterative proposal-rejection algorithm in which pairs of residuals were swapped to restore desirable statistical dependencies. At each iteration, a pair of residuals was randomly selected, and the ratios

$$T_1(\boldsymbol{e}^*) = \frac{R_1(\boldsymbol{e}^*)}{R_{tot}(\boldsymbol{e}^*)}$$

and

$$T_2(\boldsymbol{e}^*) = \frac{R_2(\boldsymbol{e}^*)}{R_{tot}(\boldsymbol{e}^*)}$$

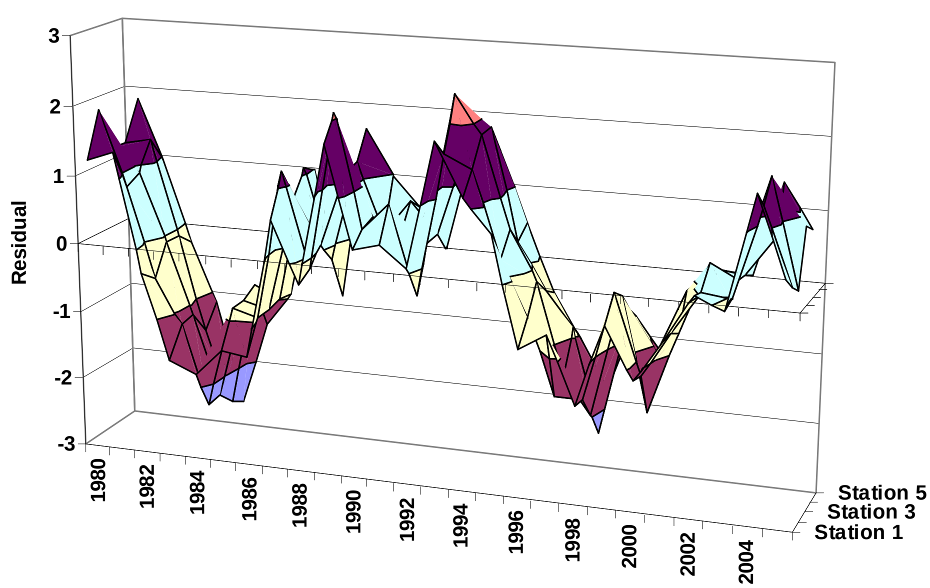
were computed before and after swapping the selected pair. If the swap decreased the Euclidean distance

$$\sqrt{(T_1(\boldsymbol{e}^*) - T_1(\boldsymbol{e}))^2 + (T_2(\boldsymbol{e}^*) - T_2(\boldsymbol{e}))^2}$$

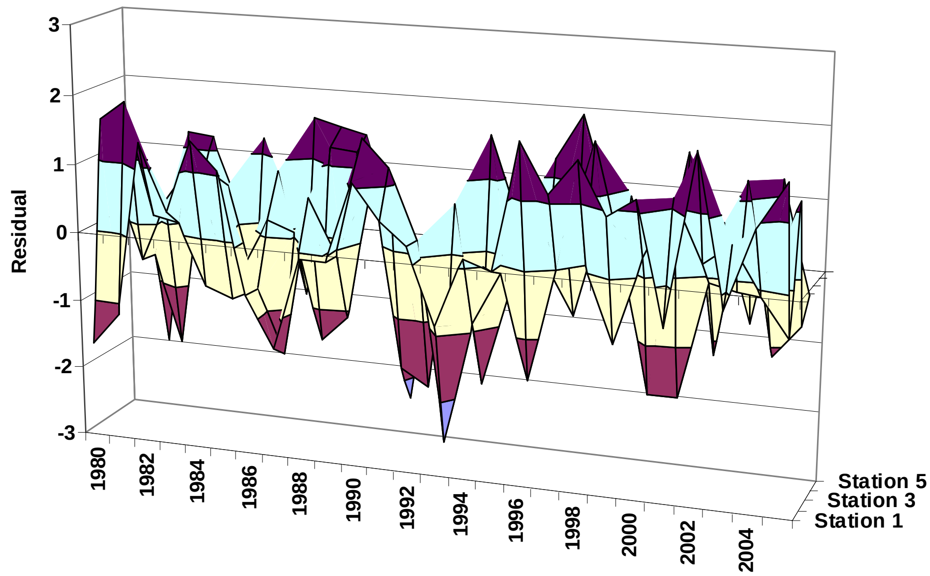
to the target it was accepted, otherwise it was rejected. The swapping was stopped after a predefined number of proposed swaps or consecutive rejections.

Figures 3.2 and 3.3 illustrate how the swapping can modify the statistical features of the bootstrap residuals. The upper graph in the first figure shows residuals with a strong correlation across series. The ordinary bootstrap subsequently destroyed that correlation, but it was restored by the swapping. The upper graph in the second figure demonstrates statistically independent residuals. Ordinary bootstrap then produced a new set of statistically independent residuals, and that configuration was left almost unchanged during the swapping.

a)



b)



c)

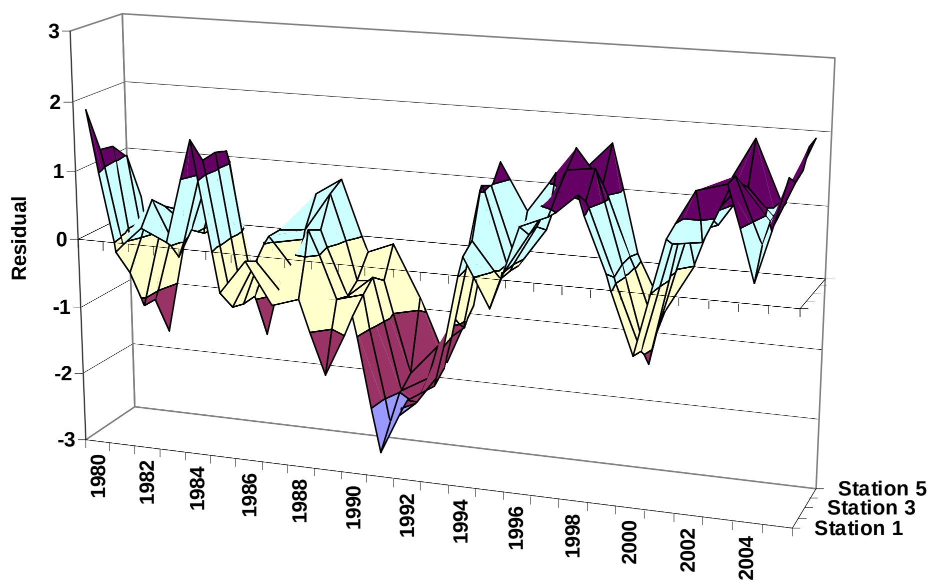
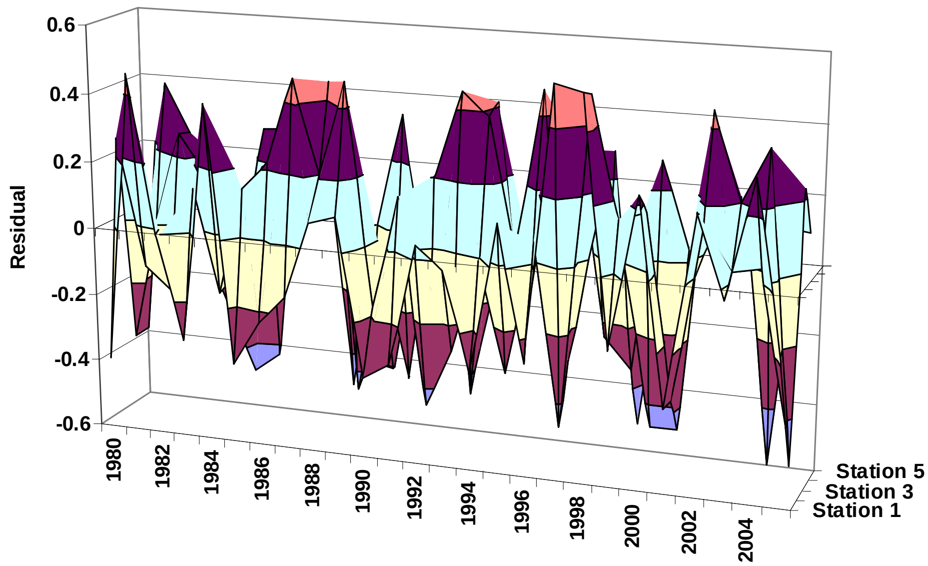
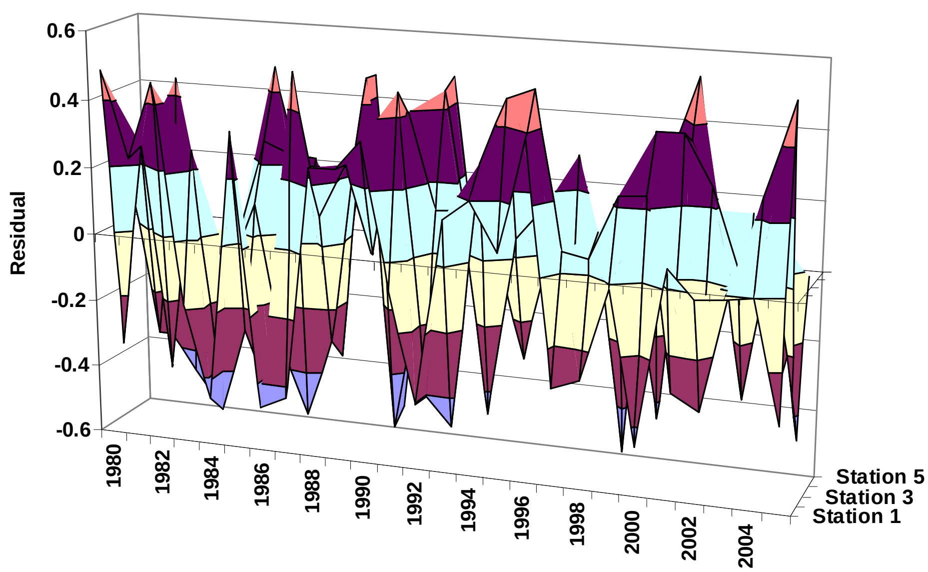


Figure 3.2 Constrained resampling of residuals that are strongly correlated across stations. The three graphs show the original data (a), an ordinary bootstrap sample of that dataset (b), and the same bootstrap sample after 100,000 proposed swaps (c).

a)


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Residual
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
Station 5
Station 3
Station 1

b)


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
Residual
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
Station 5
Station 3
Station 1

c)

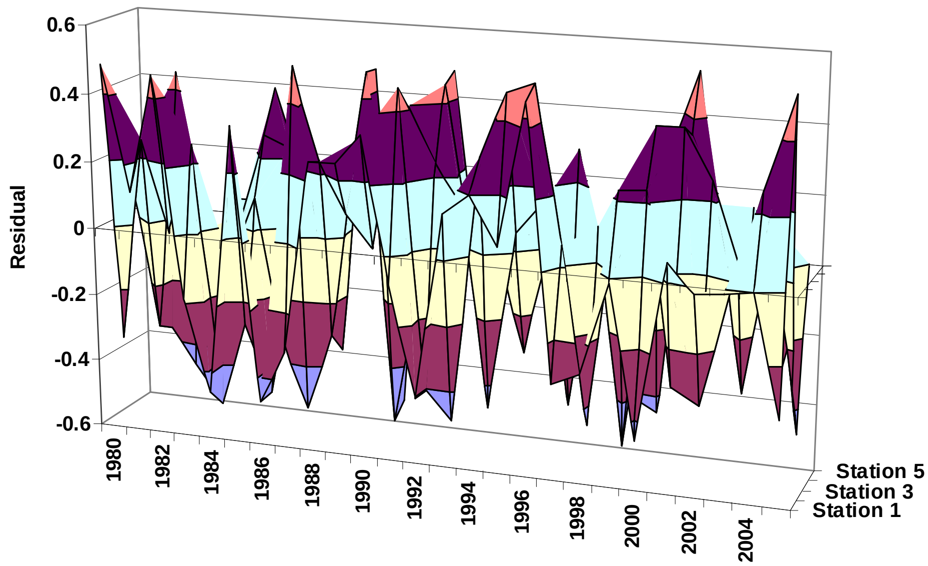


Figure 3.3 Constrained resampling of statistically independent residuals. The three graphs show the original data (a), an ordinary bootstrap sample of that dataset (b), and the same bootstrap sample after 100,000 proposed swaps (c).

The resampling procedure just presented was designed for the case of one observation per cell. For the general case with a varying number of observations per cell, we introduced the random effect model

$$e_{t,k}^{(j)} = \delta_t^{(j)} + \eta_{t,k}^{(j)}, \quad t = 1, ..., n, \quad j = 1, ..., m, \quad k = 1, ..., n(t, j)$$

where $e_{t,k}^{(j)}$ denotes the kth residual of the jth series at time t. First, the cell-specific random effects $\delta_t^{(j)}$ were predicted using expressions of the form

$$\hat{\delta}_t^{(j)} = \rho_t^{(j)} \bar{e}_{t.}^{(j)}$$

where

$$\rho_t^{(j)} = \frac{\hat{V}(\boldsymbol{\delta})}{\hat{V}(\boldsymbol{\delta}) + \hat{V}(\boldsymbol{\eta}) / n(t, j)}$$

and $\hat{V}(\boldsymbol{\delta})$ and $\hat{V}(\boldsymbol{\eta})$ denote the estimated variances of the two types of random effects (Hall & Maiti, 2006). Thereafter the $\eta_{t,k}^{(j)}$ were predicted by subtracting the predicted cell-specific components from the original residuals. Finally, constrained resampling was used to sample the $\hat{\delta}_t^{(j)}$, whereas the ordinary bootstrap was applied to the $\hat{\eta}_{t,k}^{(j)}$ representing variation within cells.

In Multitrend, we normally repeated the resampling 200 times and computed the empirical standard deviation of the estimators under consideration. Furthermore, we examined the sample standard deviation of the resampled residuals and compared it with the sample standard deviation of the residuals obtained when our model was fitted to resampled data. The ratio of the two sample standard deviations was used to adjust the empirical standard deviations for the degrees of freedom of the semiparametric regression model.

3.3 Determination of smoothing factors using block cross-validation

Cross-validation is a widely used technique to estimate the predictive power of a model and to select models of suitable complexity by splitting the entire dataset into training sets and test sets. In its simplest form, cross-validation comprises three steps (Shao, 1993): (i) the model is fitted to a training set; (ii) the fitted model is used to predict the observed responses in a test set containing all remaining data; (iii) the prediction error sum of squares (*press*) is computed for the selected test set. *Block cross-validation* refers to methods in which the original dataset is split into non-overlapping blocks, and the three steps listed above are repeated for all test sets consisting of data from a single block. Leave-one-out cross validation refers to block size one.

In this thesis, we used block cross-validation to select suitable smoothing factors λ_1 and λ_2 in the semiparametric regression model. The cross-validation was repeated for different levels of these smoothing factors, and a simple search algorithm was employed to determine the levels that maximized the predictive power of the model. Blocks were formed by joining all observations made the same year. A simulation study has indicated that such blocks represent a reasonable compromise between statistical efficiency and robustness to correlation among the observed responses (Libiseller & Grimvall, 2003).

3.4 Normalization with respect to covariates

Environmental data often exhibit substantial natural variability caused by the weather conditions at or prior to the sampling occasion. Our semiparametric regression model enables normalization (or adjustment) of observed responses for the levels of a set of user-defined covariates. For example, the observed concentrations of substances in river water samples can be normalized to an average runoff. If successful, such operations can remove or reduce irrelevant variation in the collected data and thereby also clarify the impact of human interventions on the environment. Meteorological normalization and other statistical adjustments are often used for trend assessment of environmental quality, in particular when considering data on air and water (e.g., Clark *et al*., 2000; Hussian *et al*., 2004; Libiseller & Grimvall, 2003; Thompson *et al*., 2001). The normalization formula in our semiparametric model can be written

$$\tilde{y}_t^{(j)} = y_t^{(j)} - \sum_{i=1}^{p} \hat{\beta}_i^{(j)} \left(x_{it}^{(j)} - \bar{x}_{i.}^{(j)} \right), \quad j = 1, \ldots, m, \quad t = 1, \ldots, n$$

and further details can be found in previous publications from our group (Hussian *et al*., 2004; Libiseller & Grimvall, 2003; Stålnacke & Grimvall, 2001).

Figure 3.4 illustrates how the influence of variability in runoff and amount of particulate matter was removed from observed flow-weighted mean concentrations of total phosphorus in the Ångermanälven River, which discharges into the Bothnian Sea.

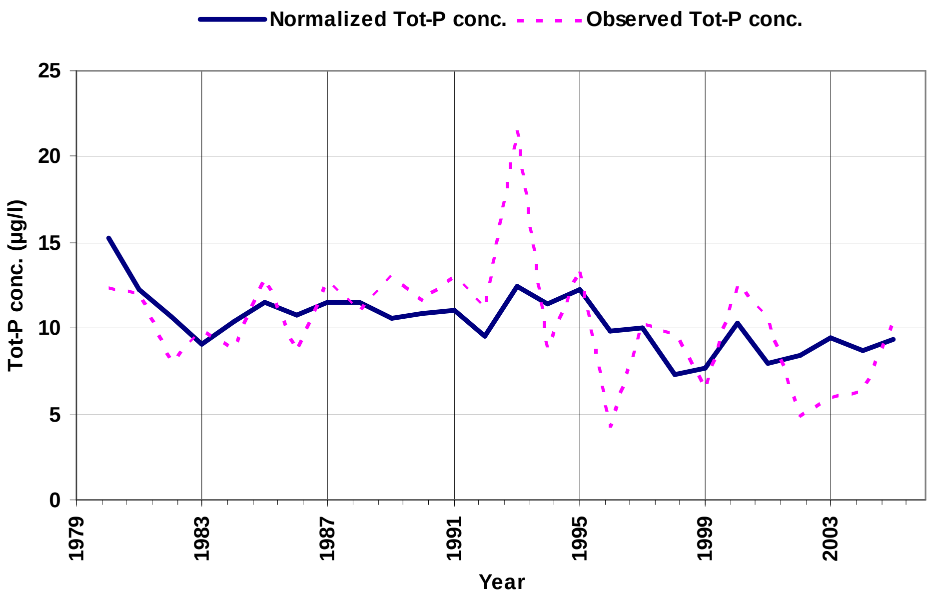


Figure 3.4 Flow-weighted annual mean concentrations of total phosphorus (*Tot-P*) in the Ångermanälven River before or after normalization with respect to runoff and amount of particulate matter.

4 Change-point detection

In climate research, a time series of observational data is said to be *homogeneous* if its temporal variation is caused solely by variations in weather and climate. However, many such series are contaminated with artificial level shifts, which can, for example, be due to modifications in measurement devices, relocation of sampling sites, or changes in the immediate vicinity of the measurement station. To enable correct interpretation of observed trends, it is obviously necessary to develop efficient techniques to detect change points in and homogenization of the collected data series. Climatologists have played a leading role in that context (Peterson *et al*., 1998).

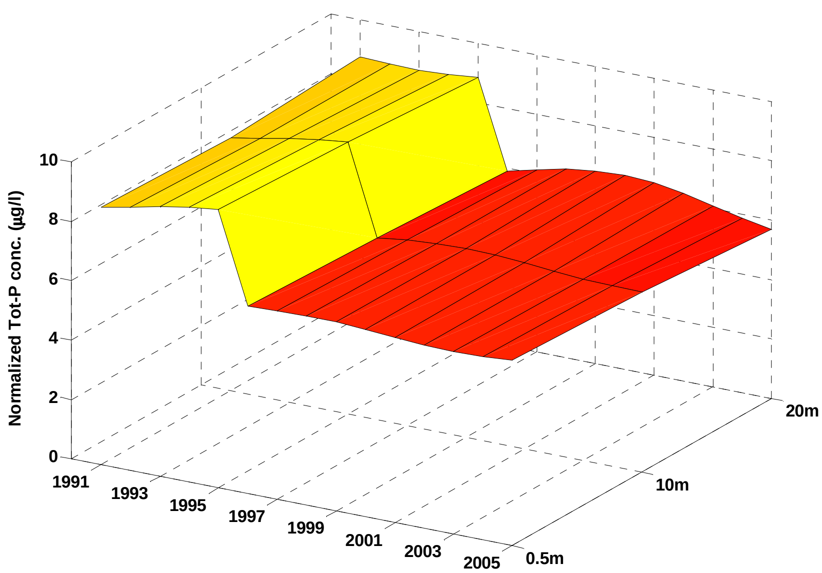


Figure 4.1 Smooth trend surface augmented with a discontinuity between 1995 and 1996 fitted to total phosphorus (*Tot-P*) levels in surface water at Dagskärsgrund in Lake Vänern. Samples were collected at depths of 0.5, 10, and 20 m.

This chapter begins by reviewing some existing methods for detecting change points and then goes on to show how the response surface model described in Chapter 3 can be augmented with a component representing abrupt level shifts. Figure 4.1 illustrates how such a model can highlight a discontinuity in observed water quality data.

4.1 Review of some existing methods

As already mentioned, climate research has been an important driving force in the development of methods for detection of change points and homogenization of long time series. In general, a climate series shows substantial short-term variability but only a weak or very weak long-term trend. If two time series represent the same region, and the difference between the two series is computed, it might even be assumed that the climate signal will be practically eliminated. Accordingly, change-point detection is often based on models in which the mean is constant if there are no artificial level shifts.

The presence of a shift in the mean of a normal distribution at some unknown instant can be tested with a likelihood ratio test (Hawkins, 1977; Worsley, 1979), and a multivariate extension of that test is also available (Sristava & Worsley, 1986). Scientists have embedded the cited tests in procedures in which the climate signal of a candidate series is removed by subtracting a reference series. Furthermore, they have addressed the problem of detecting an unknown number of change points in multiple time series. Some of the major achievements are described in the following sections.

4.1.1 The standard normal homogeneity test for a single change point

The standard normal homogeneity test (SNHT) is a likelihood ratio test proposed by Alexandersson (1986). It is applied to time series data $\{z_1, \ldots, z_n\}$, which are obtained by first subtracting a homogenous reference series from the time series to be scrutinized and then standardizing the series of differences to mean zero and variance one. The null hypothesis in the SNHT implies that the mean is zero for all time points, and the alternative hypothesis involves a single shift in the mean. The test statistic is the maximum of

$$T_v = v(\bar{z}_1)^2 + (n-v)(\bar{z}_2)^2$$

where $\bar{z}_1$ is the mean of z_1 to z_v, $\bar{z}_2$ is the mean of z_{v+1} to z_n, and $1 \le v \le n$.

4.1.2 A fixed-effect model for simultaneous detection of multiple change points

Caussinus and Mestre (2004) have developed a procedure to determine an unknown number of change points in a vector time series. For a given set of change points the measured values are given by a linear model with fixed effects. One group of these effects defines the mean response and the artificial level shifts in the measured data. Another group represents annual weather effects common to all stations in the investigated region. A stopping rule determines the number of change points.

To enable a more precise definition of the outlined procedure, let the vector

$$\boldsymbol{Y} = (Y_1^{(1)}, ..., Y_1^{(m)}, ..., Y_n^{(1)}, ..., Y_n^{(m)})^T$$

denote observations made at m stations during a period of n years, and let $u = (u_1, ..., u_n)^T$ be a vector of annual weather effects. Further, let K_1, …, K_m indicate the number of segments defined by the change points in each of the m data series, and let the vector.

$$\boldsymbol{\mu} = (\mu_1^{(1)}, ..., \mu_{K_1}^{(1)}, ..., \mu_1^{(m)}, ..., \mu_{K_m}^{(m)})^T$$

represent segment-specific effects. Then, the vector of observed responses can be written on the form

$$\boldsymbol{Y} = \boldsymbol{S\mu} + \boldsymbol{Tu} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{\varepsilon} = (\varepsilon_1^{(1)}, ..., \varepsilon_1^{(m)}, ..., \varepsilon_n^{(1)}, ..., \varepsilon_n^{(m)})^T$$

is a vector of noise components, and ***S*** and ***T*** are incidence matrices in which each row consists of zeros and a single one. The role of ***S*** is to assign a segment to each observation, and ***T*** is used to indicate the year of each observation. More specifically, the two matrices can be written

$$\boldsymbol{S} = \begin{pmatrix} S_1^{(1)} & 0 & . & 0 \\ 0 & S_1^{(2)} & . & . \\ . & . & . & 0 \\ 0 & . & 0 & S_1^{(m)} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ S_n^{(1)} & 0 & . & 0 \\ 0 & S_n^{(2)} & . & . \\ . & . & . & 0 \\ 0 & . & 0 & S_n^{(m)} \end{pmatrix} \text{ and } \boldsymbol{T} = \begin{pmatrix} 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & 0 & 1 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & 0 & 1 \end{pmatrix}$$

where $S_t^{(j)}$ is a vector of length K_j with K_j-1 zeros and a single one indicating the segment of the *j*th series to which $Y_t^{(j)}$ belongs. The constraint

$$\overline{u} = \frac{(u_1 + ... + u_n)}{n} = 0$$

is introduced to make the parameters identifiable.

For a given number of change points $K = K_1 + \ldots + K_m$, the model parameters are estimated using a least squares algorithm in which the optimal combination of time points and sizes of the level shifts is determined. The number of change points can be estimated using a penalized likelihood approach. Letting $\hat{E}(Y_t^{(j)}(K))$ denote the least square estimate of $E(Y_t^{(j)}(K))$ in the optimal model with a total of K change points and $\hat{E}_t^{(j)}(0)$ the least square estimate of $E_t^{(j)}(0)$ in a model without change points

$$C_K(Y) = \ln\left[1 - \frac{\sum_{j=1}^{m}\sum_{t=1}^{n}\left\{\left(\hat{E}(Y_t^{(j)}(K)\right)^2 - \left(\hat{E}_t^{(j)}(0)\right)^2\right\}}{\sum_{j=1}^{m}\sum_{t=1}^{n}\left\{Y_t^{(j)} - \hat{E}_t^{(j)}(0)\right\}^2}\right] + \frac{2K}{nm-m-n+1}\ln(nm)$$

is computed for all K, and K^* = $\arg\min_K\{C_K(Y)\}$ is the estimated number of change points.

4.1.3 A mixed linear model for sequential detection of change points

Another method used to detect multiple change points is designated MASH (multiple analysis of series for homogenization; Szentimrey, 2000). The underlying probability model is a mixed linear model, which implies that, in contrast to Caussinus and Mestre's model, MASH can take into account the covariance structure of data representing different stations. On the other hand, it does not include any procedure to simultaneously estimate all change points.

In matrix form, the model behind MASH can be written

$$\boldsymbol{Y} = \boldsymbol{S\mu} + \boldsymbol{TU} + \boldsymbol{\varepsilon}$$

where $U = (U_1, ..., U_n)^T$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_{nm})^T$ denote zero mean Gaussian random vectors, $\boldsymbol{\mu}$ is a vector of segment-specific fixed effects, and $\boldsymbol{S}$ and $\boldsymbol{T}$ play the same role as mentioned in section 4.1.2. In general, the coordinates of the $\boldsymbol{\varepsilon}$-vector are assumed to be independent, whereas $\boldsymbol{U}$ can have an arbitrary covariance matrix.

The parameter estimation is based on a procedure in which an optimal reference is constructed for each candidate series. More specifically, each reference series is a weighted sum of all series other than the candidate series, and the weights are selected so that the variance in the candidate-reference differences is minimized. Significance tests are utilized to determine the number of change points.

4.1.4 A general procedure for simultaneous detection of multiple change points

Picard and co-workers (2007) have developed a general segmentation method for a sequence of Gaussian vectors with an unknown number of level shifts. The number of segments and their length can differ from coordinate to coordinate, but the mean is always constant in each segment. In addition, it is worth noting that the cited method seems to combine important advantages of MASH and the procedure suggested by Caussinus and Mestre. Like MASH, Picard's technique is based on a linear model

$$\boldsymbol{Y} = \boldsymbol{S\mu} + \boldsymbol{TU} + \boldsymbol{\varepsilon}$$

with both fixed and random effects, and, as in Caussinus and Mestre's method, all change points are estimated simultaneously. The latter is achieved by using an expectation-maximization (EM) algorithm (Hastie *et al.*, 2001) to estimate all model parameters, including the covariance matrix of $\boldsymbol{Y}$.

4.2 Detection of change points in the presence of smooth trends

When the trends in the different coordinates of a vector time series are significantly different, the above-mentioned methods with piecewise constant means are not appropriate. In particular, there is a need for techniques that are easy to comprehend, yet capable of detecting change points in the presence of smooth trends that may differ from coordinate to coordinate. In the current work, we developed such methods by extending the response surface models presented in Chapter 3. More specifically, we introduced models of the form

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

where $\boldsymbol{\gamma}_t = \left(\gamma_t^{(1)}, \ldots, \gamma_t^{(m)}\right)^T$, $t = 1, \ldots, n$ defines contemporaneous abrupt level shifts in all the investigated series.

The detection and estimation of artificial level shifts was based on several different adjustment functions γ_t, t = 1, …, n. In the case of a single level shift that occurred simultaneously at all stations, we used adjustment functions of the form

$$\gamma_t^{(j)} = \begin{cases} \mu, \text{if } t \le t_1 \\ \mu + \theta(j), \text{if } t > t_1 \end{cases}$$

where $\theta(j)$ denotes the level shift in the jth coordinate between times t_1 and t_1+1. Furthermore, we introduced simple parameterizations, such as

$$\theta(j) = \theta_0, \quad j = 1, ..., m$$

or

$$\theta(j) = \theta_0 + \theta_1 j, \quad j = 1, ..., m$$

Because the parameter μ is not identifiable in the presence of α, it was normally selected so that the average adjustment was zero.

Adjustment functions involving two change points were defined analogously. In particular, we used the parameterization

$$\gamma_t^{(j)} = \begin{cases} \mu, \text{if } t \le t_1 \text{ or } t > t_2 \\ \mu + \theta(j), \text{if } t_1 < t \le t_2 \end{cases}$$

when we searched for a period when the measured values were biased. Moreover, when a change point occurred in the middle of a year and, hence, influenced two consecutive response vectors, we estimated functions of the form

$$\gamma_t^{(j)} = \begin{cases} \mu, \text{if } t \le t_1 \\ \mu + \delta\theta(j), \text{if } t = t_1 + 1 \\ \mu + \theta(j), \text{if } t > t_1 + 1 \end{cases}$$

where $0 < \delta < 1$.

In our software Multitrend, the parameters were estimated by using a back-fitting algorithm in which estimation of α alternated with estimation of β and γ. Further details are given in Paper IV.

5 Assessing data quality and trends in surface water records

Time series representing a network of stations are often analysed one by one, but the work reported in this chapter demonstrated the added value of performing a joint analysis of multiple data series. In particular, it was noted how synchronous increases and decreases in water quality can be revealed without utilizing complex space-time modelling. The response surface methodology employed has already been described in Chapters 3 and 4.

The aim at the onset of this study was to extract major long-term trends from a water quality database that was considered to be of high quality. However, the results obtained made it necessary to focus on data quality issues and detection of artificial level shifts. Substantial parts of the results provided by the methods discussed in Chapter 3 were published in Paper II, and our analysis of change points presented in the following sections was based on the information given in Paper IV.

5.1 Investigated data

Concentration data from major rivers and the two largest lakes in Sweden were acquired from the Swedish University of Agricultural Sciences (SLU, 2008), and runoff data were provided by the Swedish Meteorological and Hydrological Institute. Figure 5.1 and Tables 5.1 and 5.2 contain information about the sampling sites and water quality parameters that were investigated.

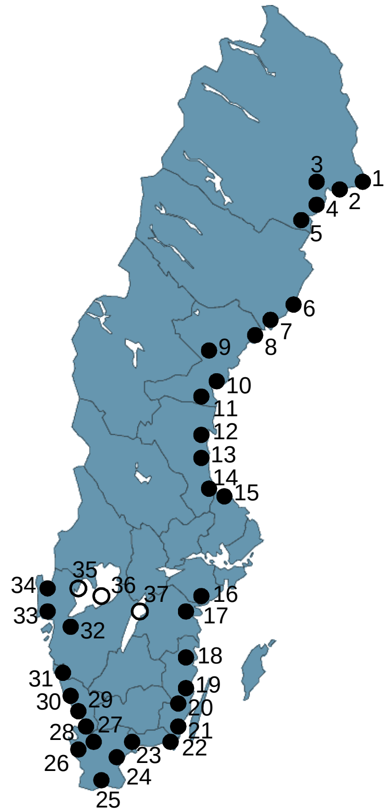


Figure 5.1 Map of Sweden with location of sampling sites in the investigated rivers and lakes.

Table 5.1 Water quality parameters and number of observations for the investigated water bodies

Water quality parameter	Time span	No. of observations*
Total nitrogen (persulphate digestion)	1988–2005	200–222
Kjeldahl nitrogen	1980–2005	282–312
Sum of nitrite and nitrate nitrogen	1980–2005	283–312
Ammonium nitrogen	1980–2005	283–312
Total phosphorus	1980–2005	283–312
Phosphate phosphorus	1980–2005	283–312
Total organic carbon	1987–2005	205–228
Chemical oxygen demand (permanganate consumption)	1980–2005	283–323
pH	1980–2005	283–323
Absorbance (420 nm, 25 ^{o}C, filtered and unfiltered samples)	1980–2005	283–323

*Sampling was done more often in the Skivarpsån River and less frequently in the Gideån River and Alsterån River than in the other watercourses.

Table 5.2 Recipients, sampling sites, and runoff area for the investigated rivers and sampling sites in Lakes Vänern and Vättern

Recipient	Nr	River	Sampling site	Runoff area (km^2)
Bothnian Bay	1	Torne Älv	Mattila	34,441
	2	Kalix Älv	Karlsborg	23,845
	3	Råne Älv	Niemisel	3,781
	4	Lule Älv	Luleå	25,225
	5	Pite Älv	Bölebyn	11,285
Bothnian Sea	6	Ume Älv	Stornorrfors	26,567
	7	Öre Älv	Torrböle	2,860
	8	Gide Älv	Gideåbacka	3,442
	9	Ångermanälven	Sollefteå	30,638
	10	Indalsälven	Bergeforsen	25,767
	11	Ljungan	Skallböleforsen	12,085
	12	Delångersån	Iggesund	1,992
	13	Ljusnan	Ljusne Strömmar	19,820
	14	Gavleån	Gävle	2,453
	15	Dalälven	Älvkarleby	28,921
Southern Baltic	16	Nyköpingsån	Spånga	3,589
	17	Motala Ström	Norrköping	15,387
	18	Botorpström	Brunnsö	975
	19	Emån	Emsfors	4,441
	20	Alsterån	Getebro	1,333
	21	Ljungbyån	Ljungbyholm	735
	22	Lyckebyån	Lyckeby	810
	23	Mörrumsån	Mörrum	3,365
	24	Helgeån	Hammarsjön	4,144
	25	Skivarpsån	Skivarp	102
	26	Råån	Helsingborg	166
Kattegat and Skagerrak	27	Rönneån	Klippan	963
	28	Lagan	Laholm	6,133
	29	Nissan	Halmstad	2,677
	30	Ätran	Falkenberg	3,340
	31	Viskan	Åsbro	2,160
	32	Göta Älv	Trollhättan	47,035
	33	Örekilsälven	Munkedal	1,335
	34	Enningdalsälven	N. Bullaren	631
Lake Vänern	35		Megrundet	
	36		Dagskärsgrund	
Lake Vättern	37		Jungfrun	

5.2 Examples of synchronous temporal changes

5.2.1 Total phosphorus

Simple scatter charts of observed phosphorus concentrations versus time indicated that, in the majority of the investigated rivers, the concentrations decreased from 1980 to 1983 and then increased, and there also seemed to be a drop around 1996. This temporal pattern emerged more clearly when the response surface methodology presented in Chapter 3 was used to fit smooth trend surfaces to annual normalized concentrations for selected groups of rivers. Figure 5.2 illustrates the results obtained for fifteen rivers flowing into the Bothnian Bay and Bothnian Sea when the observed levels of total phosphorus were normalized with respect to water discharge and amount of particulate matter (the latter measured as the difference in absorbance between unfiltered and filtered water samples).

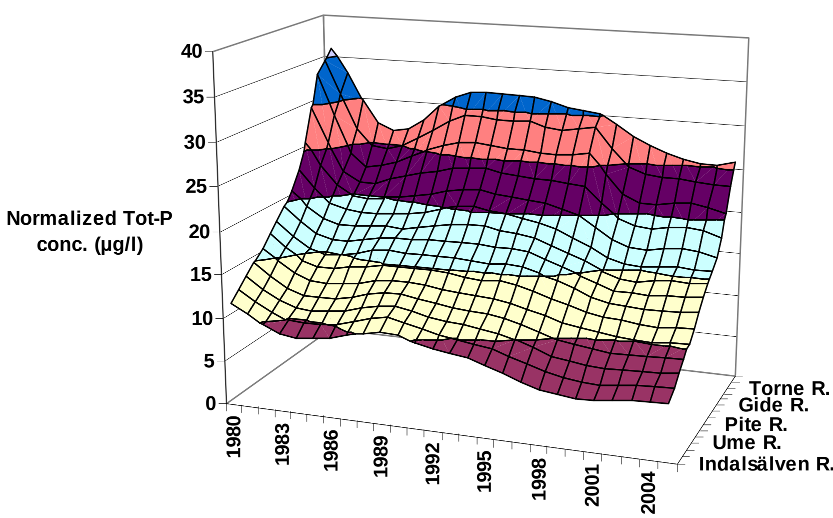


Figure 5.2 Trend surface fitted to normalized annual summaries of total phosphorus (*Tot-P*) concentrations in fifteen Swedish rivers flowing into the Bothnian Bay and Bothnian Sea (Table 5.2). The normalization was done with respect to water discharge and the amount of particulate matter, measured as the difference in absorbance between unfiltered and filtered samples.

Closer examination of the data underlying Figure 5.2 indicated that the synchronous decrease around 1996 actually was a step change. This discontinuity emerged even more clearly when the analysis was restricted to the four rivers that had the lowest frequency of outliers. Because a change in laboratory practice took place in the middle of 1996, we used a model in which the discontinuity was split between two consecutive years. Furthermore, we used water discharge as a covariate and allowed the size of the discontinuity to vary with the average phosphorus concentration in the analysed river (see Paper IV). Figure 5.3 shows the fitted trend surface, with the detected discontinuity, for those four rivers and Table 5.3 presents the estimated level shifts and their standard errors. In particular, it can be noted that level shifts also occurred in rivers where the measured phosphorus concentrations were far above the limit of detection of the analytical procedure.

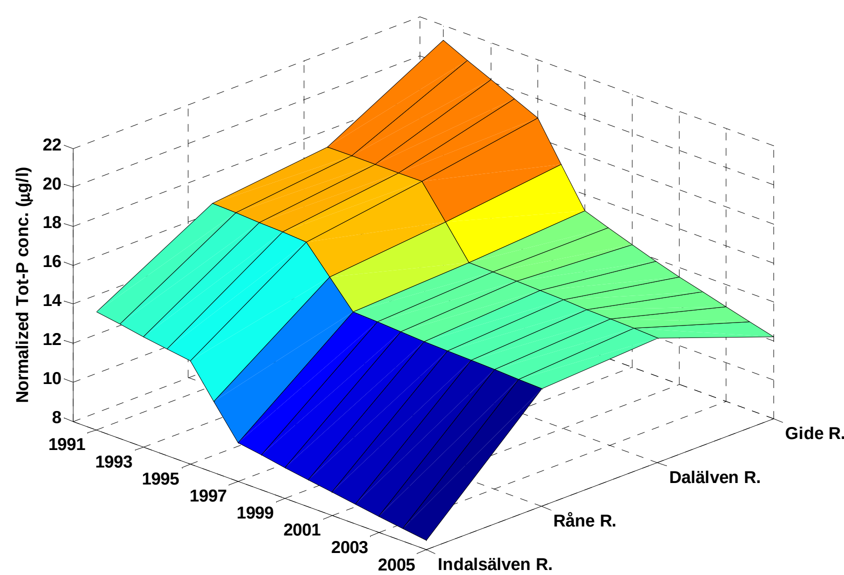


Figure 5.3 Trend surface with discontinuities fitted to total phosphorus (*Tot-P*) levels in four major rivers in northern Sweden. The statistical model and the sampled rivers were the same as in Table 5.3.

Table 5.3 Estimated level shifts in total phosphorus concentrations recorded in four rivers in northern Sweden. The model had level shifts that were equally split between 1995–1996 and 1996–1997, and the size of the level shifts was allowed to vary with the sampled river

River	Level shift (μg/l)	Standard error (μg/l)
Indalsälven	–2.90927	1.130746
Råne	–2.61134	0.774648
Dalälven	–3.26740	1.115243
Gide	–2.89887	1.348316
Average	–2.92172	0.875484

In search of further evidence of synchronous changes in phosphorus concentrations, we also analysed data from Lakes Vänern and Vättern (Paper II). Figure 5.4 shows that the temporal changes in phosphorus concentrations at Dagskärsgrund in Lake Vänern were almost identical at depths ranging from 0.5 to 20 m. In addition, the trough in the early 1980s and the drop around 1996 that we had found in the data from rivers in northern Sweden were also observed here.

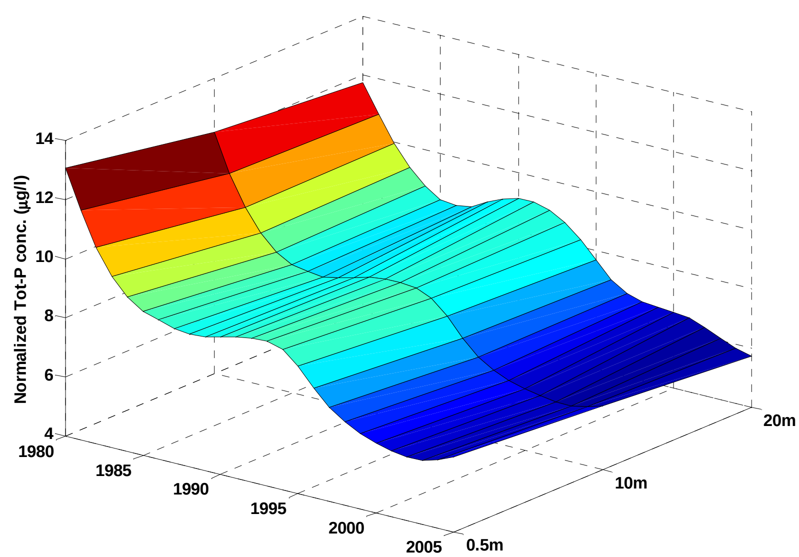


Figure 5.4 Trend surface fitted to temperature-normalized concentrations of total phosphorus (*Tot-P*) at Dagskärsgrund in Lake Vänern. The samples included in this analysis were collected at three different depths (0.5, 10, and 20 m).

Because the response surface methodology used to obtain the diagram in Figure 5.4 may have smoothed out step changes to produce gradual changes, we also plotted the individual concentration records along with trends fitted separately to data from 1980–1995 and 1996–2005 (Figure 5.5). This graph strongly indicated that the decrease around 1996 was actually a step change, whereas the decrease in the early 1980s seemed to include changes spread out over several years.

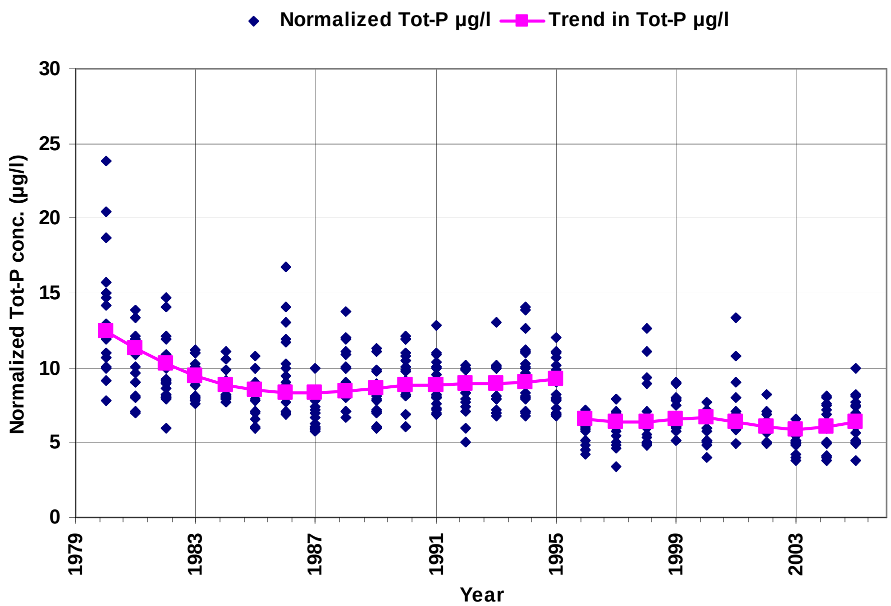


Figure 5.5 Trend curve for temperature-normalized concentrations of total phosphorus (*Tot-P*) shown along with individual concentration records of that element in samples collected at different depths (0.5, 10, and 20 m) at Dagskärsgrund in Lake Vänern.

To quantify the step change between 1995 and 1996, we fitted the change point model described in Chapter 4 to data from 1991 and onwards, and the results are illustrated in Figure 5.6. The size of the discontinuity was estimated to 3.1 µg/l, and residual resampling showed that the standard error of the estimated level shift was much smaller (0.44 µg/l). Possible causes of these surprisingly synchronous level shifts are scrutinized in section 5.3.1.

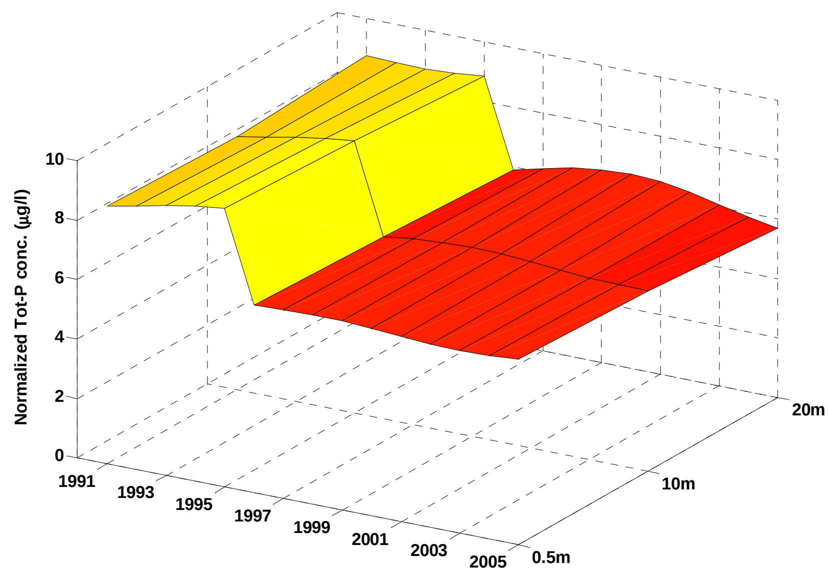


Figure 5.6 Smooth trend surface (augmented with a discontinuity between 1995 and 1996) fitted to total phosphorus (*Tot-P*) concentrations at Dagskärsgrund in Lake Vänern.

5.2.2 Total nitrogen

Statistical analyses of total nitrogen measurements based on persulphate digestion (*Tot-N*(*ps*)) revealed a pronounced downward trend that started in the mid 1990s. Figure 5.7 shows the trend observed for rivers discharging into the Kattegat and Skagerrak, and a similar pattern was found for rivers flowing into the Baltic Sea. In northern Sweden, where the average nitrogen concentration is lower, the downward trend was weaker.

Total nitrogen levels can also be determined by computing the sum of Kjeldahl nitrogen, nitrite, and nitrate nitrogen. In theory, these computed concentrations (*Tot-N*(*Kj*)) should be strongly correlated with the *Tot-N*(*ps*) values. However, simple scatter charts revealed some remarkable discrepancies. Firstly, there was a small subset of *Tot-N*(*ps*) records that indicated levels that were twice as high as they should have been, and this could be ascribed to calculation or dilution errors in the chemical analysis (Paper II). Secondly, there was an unexpectedly strong downward trend in the amount of organically bound nitrogen (*Org*-*N*(*ps*)), which was estimated by computing the difference between the *Tot*-*N*(*ps*) value and the sum of the measured concentrations of inorganic nitrogen species (ammonium, nitrite, and nitrate).

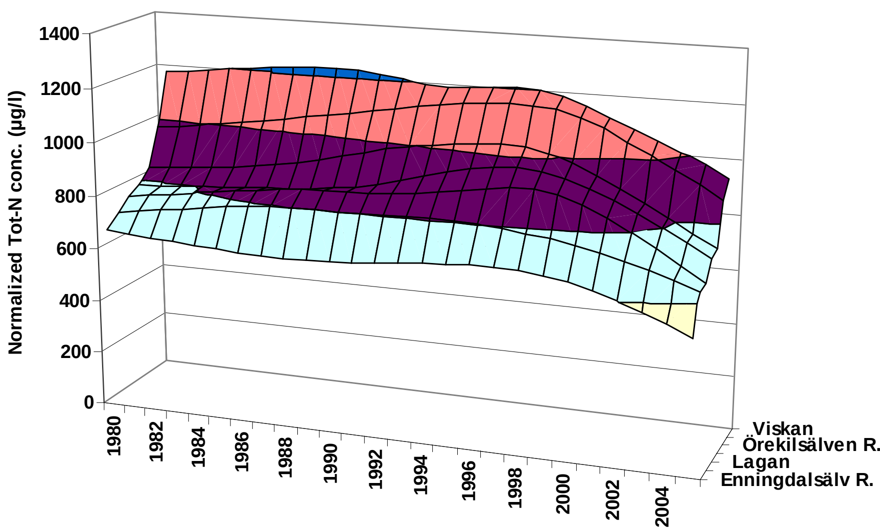


Figure 5.7 Trend surface fitted to flow-normalized concentrations of total nitrogen (*Tot-N*(*ps*)) in seven rivers (the Lagan, Nissan, Ätran, Viskan, Göta, Örekilsälven, and Enningdalsälven Rivers) discharging into the Kattegat and Skagerrak.

Figure 5.8 shows a smooth response surface fitted to *Org*-*N*(*ps*) records for the major Swedish rivers discharging into the Kattegat and Skagerrak. It seems that the computed *Org*-*N*(*ps*) values started to decline in the mid 1990s. Furthermore, this decrease was remarkably synchronous, even though the hydraulic residence times in lakes upstream of the sampling sites ranged from less than a year to almost ten years in the Göta River Basin. A similar analysis of organic nitrogen records obtained by subtracting the sum of nitrite and nitrate concentrations from Kjeldahl nitrogen did not reveal any clear temporal trends.

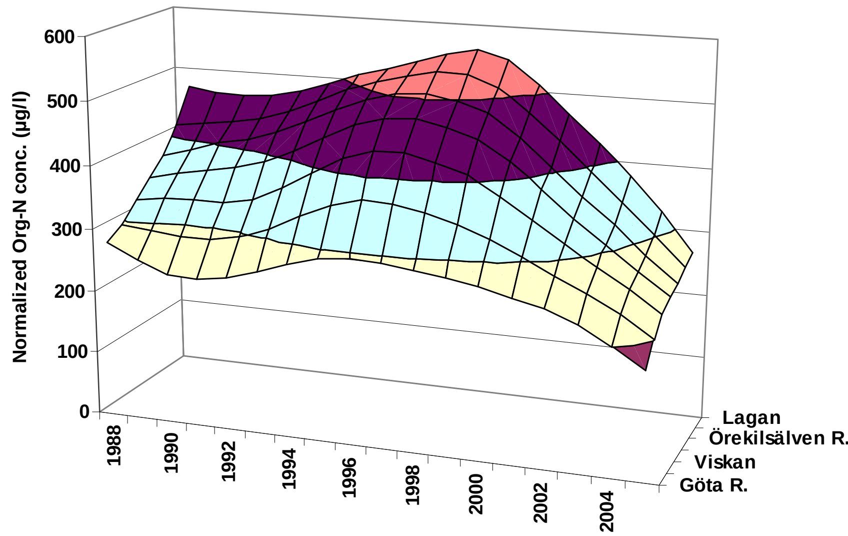


Figure 5.8 Trends in flow-normalized concentrations of organic nitrogen (*Org-N*(*ps*)) in rivers discharging into the Kattegat and Skagerrak. The investigated rivers were the same as in Figure 5.7.

Figure 5.9 provides further evidence that the choice of measure of the nitrogen content had a dramatic effect on the conclusions that can be drawn about recent temporal trends.

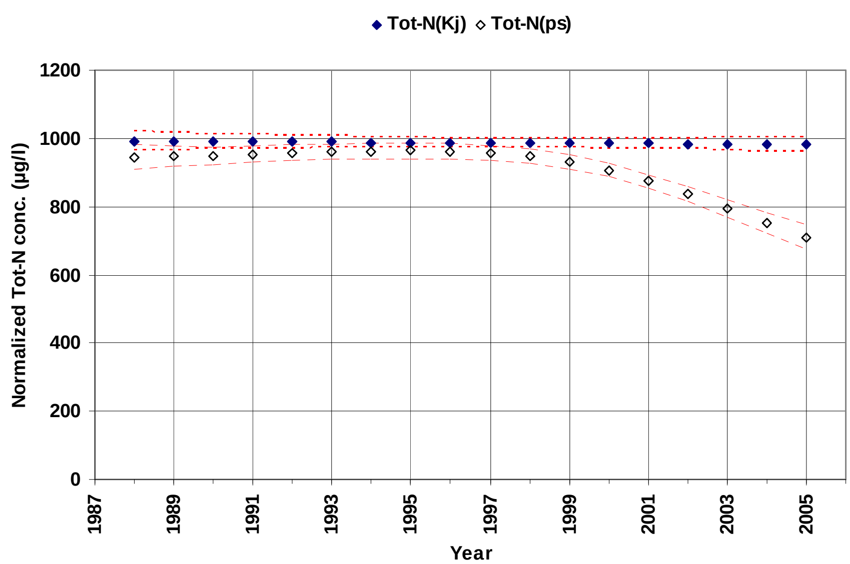


Figure 5.9 Trend lines and associated 95% confidence bands for the arithmetic mean of flow-normalized concentrations of *Tot-N*(*ps*) and *Tot-N*(*Kj*) in rivers discharging into the Kattegat and Skagerrak. The investigated rivers were the same as in Figure 5.7.

5.2.3 Organic matter

At all of the studied sites, the amount of organic carbon in the collected water samples has long been measured as chemical oxygen demand by means of potassium permanganate titration ($COD(KMnO_4)$). Since 1987, the samples have also been analysed for total organic carbon (*TOC*) using a TOC analyzer. Due to substantial interannual variation in both the TOC and COD records, it was difficult to extract any clear temporal trends from the original time series of data. However, simple plots of TOC-to-COD ratios revealed several patterns that called for further attention. Figure 5.10 shows that the values recorded in 1997 were clearly elevated, and that there was an upward tendency in the ratios calculated for 1987–1990. Moreover, almost identical temporal patterns were found when the entire dataset was split into subsets representing different regions in Sweden. Figure 5.11 shows the data for the riverine input to the Bothnian Bay and Bothnian Sea.

Using the methods described in Chapter 4 and Paper IV, we undertook an entirely data-driven search for a time period when the average TOC-to-COD-ratio differed from the general trends in the data. As expected, our algorithm identified 1997 as a period with abnormal data (see Fig. 5.12). Furthermore, the estimated level shift that year was 0.062 units, with a standard error of 0.0038.

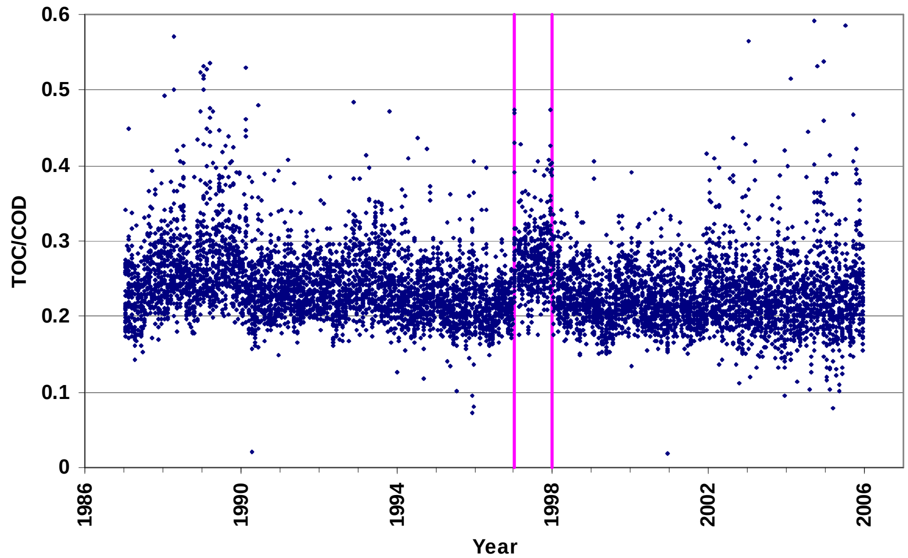


Figure 5.10 TOC-to-COD ratios calculated for all 34 of the Swedish rivers that were investigated (Table 5.2). The vertical lines indicate change points.

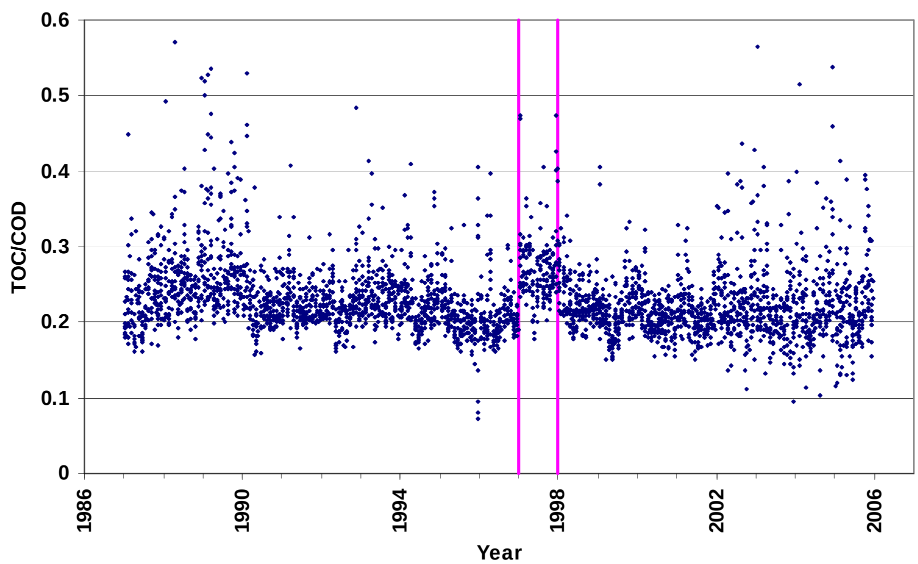


Figure 5.11 TOC-to-COD ratios calculated for fifteen rivers discharging into the Bothnian Bay and Bothnian Sea (Table 5.2). The vertical lines indicate change points.

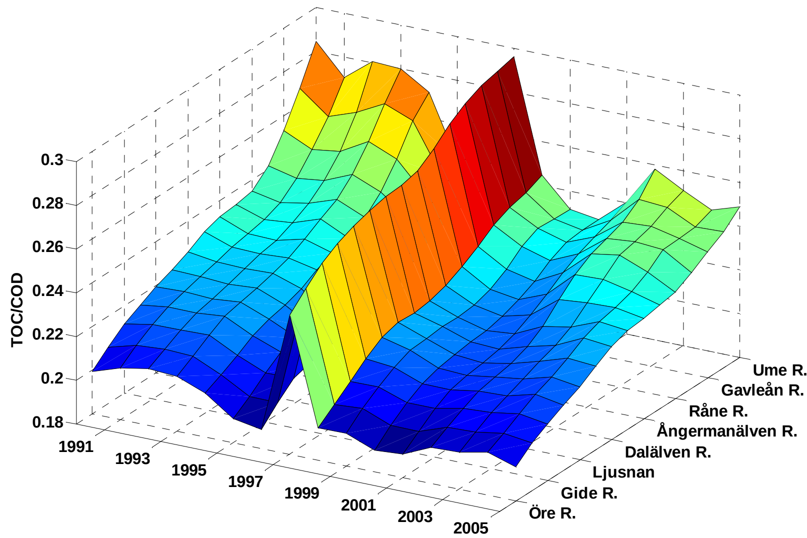


Figure 5.12 Trend surface with discontinuities fitted to the data given in Figure 5.11. Two level shifts of equal size but with different signs were assumed to be present during the period 1990–2005. The timing of the shifts was determined by an unprejudiced search.

5.3 Interpretation of observed patterns

Our study of water quality data provided numerous examples of remarkably synchronous temporal changes in rivers representing a wide range of hydrogeological conditions and anthropogenic pressures. Theoretically, there were four plausible explanations for such coinciding fluctuations: (i) large-scale human interventions; (ii) large-scale variation in weather conditions; (iii) intentional or inadvertent alterations in sampling and laboratory practices; (iv) artefacts in the statistical procedures used to analyse the collected data.

Inasmuch as the data from different regions had been analysed separately, we could rule out the fourth explanation, namely, the possibility that the remarkably synchronous changes in water quality were merely an artefact of the statistical procedures used. Moreover, we noticed that the risk of undesired smoothing across sampling sites in our response surface methodology was generally small, because the smoothing factors were selected to optimize the predictive power of the underlying regression model. In the sections that follow, the other three explanations are considered separately for each water quality parameter.

5.3.1 Total phosphorus

It is indisputable that, around 1996, the phosphorus levels decreased simultaneously in water bodies representing a wide range of anthropogenic pressures and hydraulic residence times ranging from less than a year to about 80 years. In addition, our change point analyses gave a very strong indication that the decreases that were detected in both Lake Vänern and Lake Vättern, as well as in several of the investigated rivers, had occurred rather abruptly. Considering the magnitude and spatial distribution of the observed drops in phosphorus in 1996, we ruled out the possibility that those decreases could have been largely due to anthropogenic interventions.

Internal loading triggered by specific weather conditions can occasionally cause relatively rapid changes in phosphorus concentrations in a body of water. However, inspection of water discharge and temperature data did not reveal any events that could explain why the decline in total phosphorus in 1996 was greater than all other interannual changes during the past fifteen years, and why the same pattern was found in both northern and southern Sweden in 1996. Accordingly, we also excluded the idea that the observed step change was a purely natural phenomenon.

Thus the only plausible explanation remaining comprised changes that had occurred in the sampling methods, sample handling, or chemical analyses. Notably, a report from the laboratory that conducted the chemical analyses did describe changes in the methods used to determine low concentrations of the substances of interest (Sonesten & Engblom, 2001). However, our analysis strongly indicated that the size of the artificial level shift was larger than reported by the cited authors, and also that concentrations far above the limits of detection were influenced.

Possible explanations for the trough in the early 1980s were discussed in Paper II. By the process of elimination, we concluded that this pattern in the reported data was also strongly influenced by changes in sampling or laboratory practices.

5.3.2 Total nitrogen

The dramatic decrease in the computed levels of organic nitrogen (*Org-N*(*ps*)) is an indisputable fact. Let us for a moment assume that both the total nitrogen measurements based on persulphate digestion and the observed levels of inorganic nitrogen were correct. Then there must have been an unprecedented change in the composition of organic matter in Swedish watercourses (Paper II). More precisely, an almost 50% decrease in the amount of organic nitrogen that could have been digested by persulphate would have coincided in time with a general increase in the amount of organic matter that could have been oxidized with permanganate. We regarded this as very improbable.

In search of plausible explanations for the downward trend in *Tot-N*(*ps*), it came to our knowledge that the laboratory conducting the chemical analysis had informed some clients about incomplete digestion of the organic matter in the analysed samples. Our statistical analysis of the nitrogen data revealed the magnitude and duration of those problems.

5.3.3 Organic matter

Cycles in the meteorological forcing and water pathways can be responsible for considerable temporal changes in the amount of organic matter in surface water samples, and it cannot be excluded that changes in land use may be responsible for long-term trends in such data. However, much of the variability is suppressed by calculation of TOC-to-COD ratios, and hence it is remarkable that the computed average level of those ratios suddenly increased by about 30% in 1997 and then returned to the previous level. In light of the finding that this temporary level shift occurred simultaneously in different parts of Sweden and in water bodies representing a wide range of hydrogeological conditions and anthropogenic pressures, it seemed very improbable that there were natural explanations for this abrupt change. Accordingly, we concluded that artificial level shifts had also influenced the reported concentrations of organic matter in Swedish rivers. However, we were unable to ascertain whether such artefacts had primarily affected the TOC or the COD records.

5.4 Implications for surface water monitoring

The results of our analysis of Swedish surface water data challenge the present priorities in water quality monitoring. According to Figures 5.6 and 5.9 and the associated discussion, the temporal trends in total phosphorus and total nitrogen (*Tot-N*(*ps*)) concentrations reported over the past fifteen years were influenced to a greater extent by artificial level shifts than by actual changes in the environment. Furthermore, Figure 5.12 indicates that records of organic matter were also strongly contaminated by artificial level shifts. This situation is, of course, unsatisfactory. Moreover, it is very unfortunate that the current monitoring system has incomplete information about known measurement errors, and no attempts have been made to remove systematic errors that affect large amounts of data.

Individual records that are obviously flawed can be detected and removed by methods for process control or by using more specific tools for water quality monitoring (Clement *et al.*, 2007). However, both our analysis of water quality data and the current efforts to extract temporal trends from observational climate series show that the major problems in interpreting the collected data are related to relatively small systematic errors that affect a large number of observations. Such data problems may not emerge until the whole history of measurements from an entire network of sampling sites is scrutinized. Hence, there is a strong need for a monitoring system in which conventional quality assurance is complemented with thorough statistical follow-up of reported values. Regarding the statistical tools that are needed, we noted the following:

(i) Visual inspection of scatter charts containing data from many sampling sites is an important element in any statistical evaluation of monitoring data.

(ii) Noise reduction by adjustment for covariates or formation of ratios or differences between interrelated water quality parameters can greatly facilitate the detection of trends and data quality problems.

(iii) Fitting of trend surfaces to data representing an ordered set of stations enables detection of synchronous (artificial or real) level shifts without utilizing complex space-time modelling.

(iv) Models combining smoothing and change-point detection in multiple time series of data provide a useful tool for quantifying synchronous level shifts and correcting historical data.

Visual inspection and noise reduction have long been used within the field of data analysis. However, our response surface methodologies to estimate smooth trends and change points at an ordered set of stations constitute a new and powerful tool. In addition, our new technique for resampling statistically dependent data enabled realistic calculations of confidence intervals and assessment of the statistical significance of detected level shifts.

The mentioned data quality problems also highlight the difference between an environmental information system and a monitoring system that is focused almost exclusively on data collection. Our study has demonstrated that it is possible to supplement the present raw data with filtered datasets in which obviously flawed data have been removed and other records have been adjusted for known systematic errors.

6 Assessing data quality and trends in groundwater records

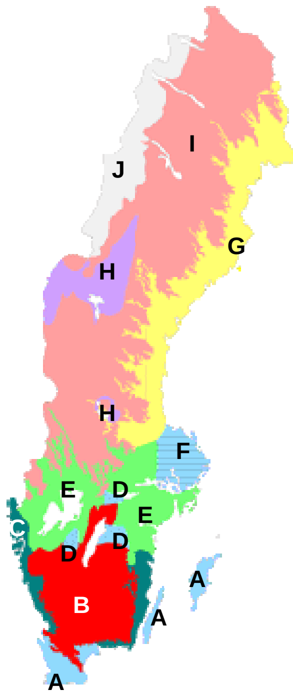
The data on groundwater quality scrutinized in the work described in this chapter proved to have a number of features that called for special attention: (i) they were relatively scattered in space and time; (ii) serial correlation of observed concentrations might have been responsible for spurious trends; (iii) the evidence of anthropogenic trends varied strongly from station to station. Accordingly, the statistical analysis focused on the problem of achieving an overview of a fairly large dataset. The results presented in this chapter are primarily based on Papers III and IV.

6.1 Investigated data

The Geological Survey of Sweden (SGU) is responsible for the national monitoring of groundwater quality. A relatively large number of groundwater bodies are normally sampled 2–6 times a year, and the physico-chemical analysis of the collected samples includes determination of major inorganic ions, pH, conductivity, and temperature (SGU, 2008). We investigated data from a total of 77 sites in ten hydrogeological regions (Fig. 6.1) where sampling has been conducted regularly at least since 1980. In particular, we examined the concentration of sulphate and the buffering capacity measured as alkalinity and acid-neutralizing capacity (ANC), and the ANC levels were computed according to

$$ANC = [Ca^{2+}] + [Mg^{2+}] + [Na^{+}] + [K^{+}] + [NH_4^{+}] - [Cl^{-}] - [SO_4^{2-}] - [NO_3^{-}]$$

It is worth noting that, since July 1992, the same laboratory has been responsible for the chemical analysis of both surface water and groundwater samples collected in the national Swedish environmental monitoring programme. Before that time, the groundwater samples were analysed at two other laboratories under contract from May 1980 to June 1984 and from July 1984 to June 1992, respectively.



Region	Number of stations
A	9
B	19
C	4
D	2
E	19
F	1
G	4
H	5
I	11
J	3
	Σ 77

Figure 6.1 Sweden divided into ten geographical regions based on bedrock, hydrology, and position relative to the highest coastline.

6.2 Detection and interpretation of trends and change points

6.2.1 Alkalinity and ANC

A search for outliers in the reported concentrations of major cations and anions revealed that a small fraction of the samples (148 out of 5,557) had at least one obviously erroneous recorded concentration, and the data on those samples were omitted from the statistical analysis. Moreover, both MK statistics for temporal trends and visual inspection of the collected data clearly indicated that local pollution, presumably from road salt, had influenced seven of the 77 sites. All data from these sites were excluded, because we were primarily interested in regional trends and the possible response to decreased acid deposition.

In an attempt to detect major patterns in the selected data, ordinary univariate MK tests were employed to examine the presence of trends in alkalinity levels. When the achieved significance levels (*p*-values) were assembled only according to hydrogeological region, there was no obvious pattern in the results that were obtained. However, after the *p*-values were sorted with respect to median alkalinity, a striking pattern emerged. As can be seen in Figure 6.2, there were significant downward trends at sites with low alkalinity, and there were significant upward trends at sites with high alkalinity. This was unexpected because (i) the acid deposition in Sweden has decreased considerably over the past two decades (Miljömål, 2008), and (ii) low alkalinity groundwaters are found primarily in aquifers with relatively short residence times. In addition, the downward trends in groundwater were contradicted by upward trends in river water. Notably, we observed the strongest upward trends in low-alkalinity rivers and sampling sites located in catchments that had previously been exposed to high sulphur deposition.

To further elucidate the existence of acidification trends in groundwater, we also examined time series of ANC levels. Figure 6.3 shows the achieved significance levels. In contrast to the results for alkalinity, the most significant upward trends in ANC were discerned for groundwaters with low to medium buffering capacity. In addition, we noted that there was generally good agreement between the ANC trends in groundwater and river water (not shown).

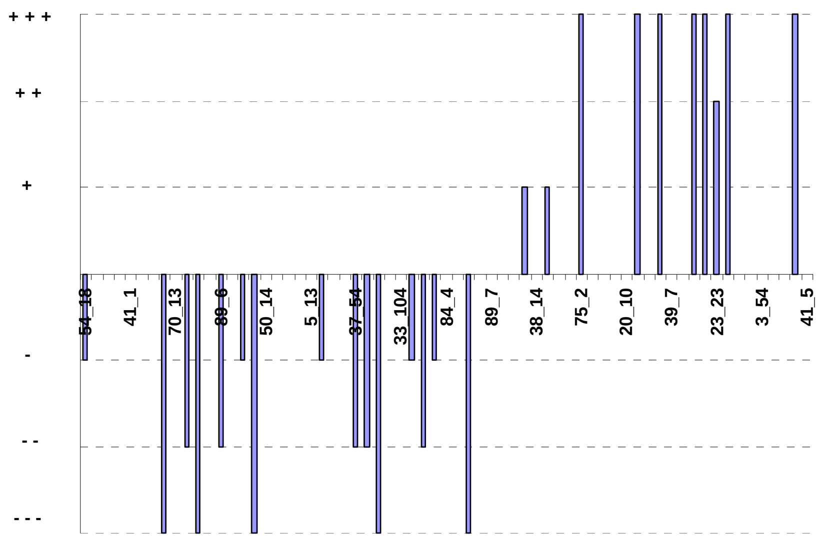


Figure 6.2 Achieved significance levels in MK tests for trends in alkalinity at 70 sites ordered according to median alkalinity. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. The station labels refer to the national Swedish groundwater monitoring programme. Three-star significances (positive and negative) were noted for (from left to right) stations 58_4, 13_107, 33_202, 19_15, 20_1, 75_2, 70_14, 3_14, 3_53, 29_8, 3_49, and 9_1.

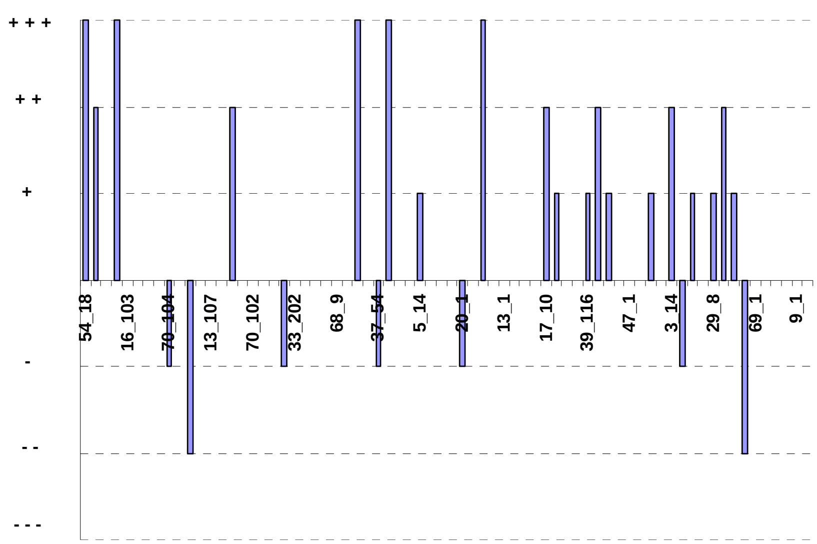


Figure 6.3 Achieved significance levels in MK tests for trends in ANC at 70 sites ordered according to median ANC. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. Three-star significances (positive) were noted for (from left to right) stations 54_18, 16_101, 37_56, 14_15, and 23_11.

Considering that both alkalinity and ANC are integrative measures of buffering capacity, we expected the two parameters to be strongly intercorrelated. However, visual inspection of scatter charts of reported alkalinity and ANC levels revealed a shift in the lowest alkalinity levels in 1984, at which time a new laboratory was engaged to conduct the chemical analysis (Paper III). After developing our new technique for change-point detection (see Chapter 4), we re-analysed the alkalinity and ANC records. More specifically, we examined the difference between alkalinity and ANC in samples with low ANC levels (less than 0.3 but greater than 0.05 meq/l). Figure 6.4 illustrates how the annual mean of the estimated trend surface (including discontinuities) was stabilized after 1984, when a new analytical procedure was introduced. Consequently, we concluded (i) that the alkalinity levels recorded during different time periods were not fully comparable, and (ii) that the ANC levels computed in the present study constituted a more reliable indicator of trends in buffering capacity.

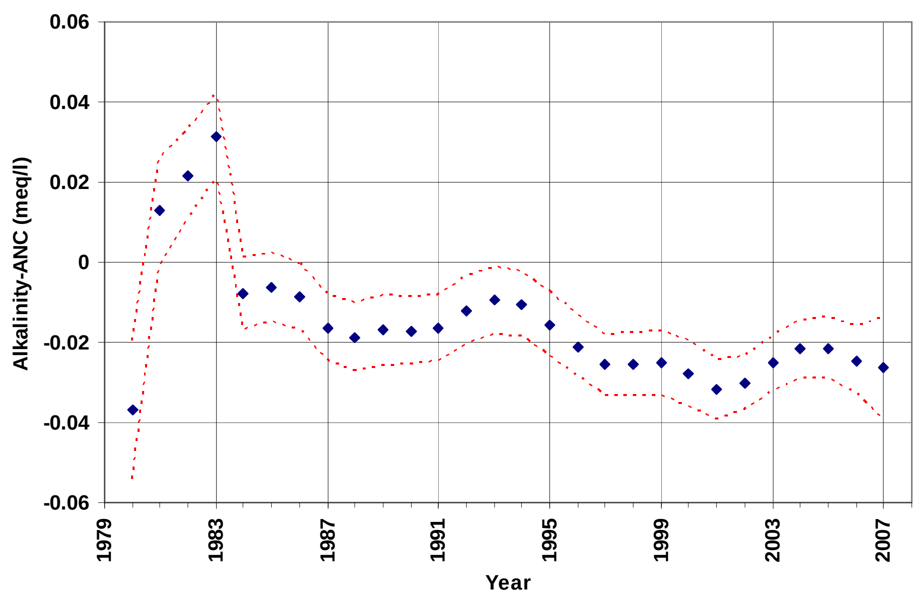


Figure 6.4 Annual means of trend levels (including discontinuities) fitted to differences between alkalinity and ANC in low-ANC samples for all groundwater stations investigated.

Further analysis of the ANC data revealed pronounced serial correlation at many of the investigated sites. Therefore, we also computed achieved significance levels in MK tests in which the effect of serial correlation was suppressed by reorganizing the data into biannual time series. However, as can be seen in Figure 6.5, there was still clear evidence of upward trends in ANC. The strongest trends prevailed in waters with low to medium alkalinity in southern Sweden, whereas the trends in northern Sweden were weak or nonexistent.

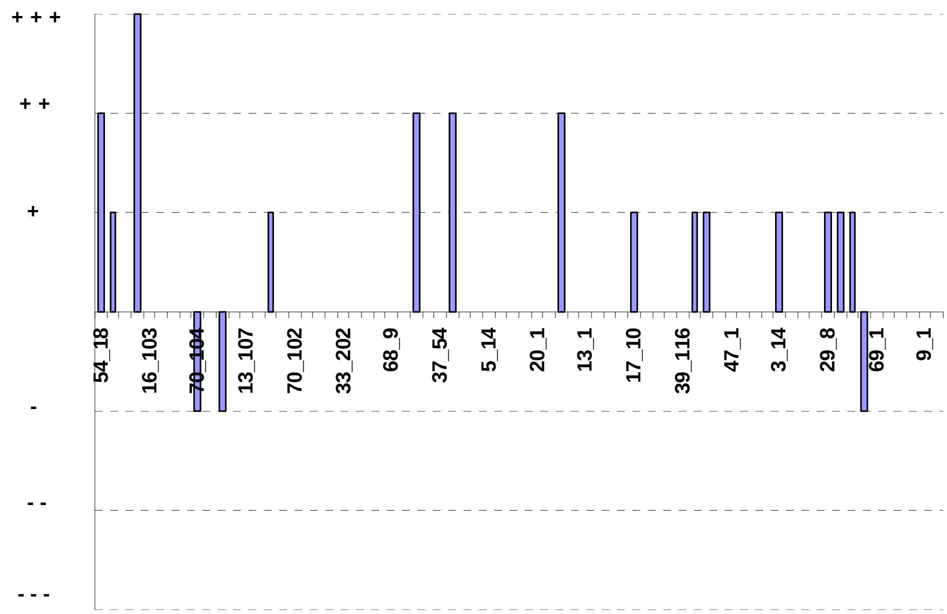


Figure 6.5 Significance in MK tests for trends in ANC at 70 sites ordered according to median ANC, showing levels achieved when the data were reorganized into time series of biannual data. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. Three-star significance (positive) was noted for station 16_101.

Chloride is sometimes used as an indicator of soil water movement, because, correctly or not (Bastviken *et al*., 2007; Schlesinger, 1997), it is considered to be inert in soil. Accordingly, we undertook partial MK tests of ANC levels, using chloride as a covariate. Furthermore, we computed ANC-to-chloride ratios that we tested for trends. Compared to the ordinary MK tests, the partial tests produced results that were almost the same, albeit slightly less significant. There were considerably fewer significant trends in the ANC-to-chloride ratios, because the formation of those ratios increased the coefficient of variation of the data that were analysed for trends.

In summary, our trend assessment provided strong evidence of upward ANC trends in the areas where acid deposition has decreased over the past decades. However, the results varied substantially between the sampling sites.

6.2.2 Potassium

Considering Figure 6.6, which illustrates potassium levels in groundwater from region B (see map in Fig. 6.1), the presence and location of discontinuities is less obvious.

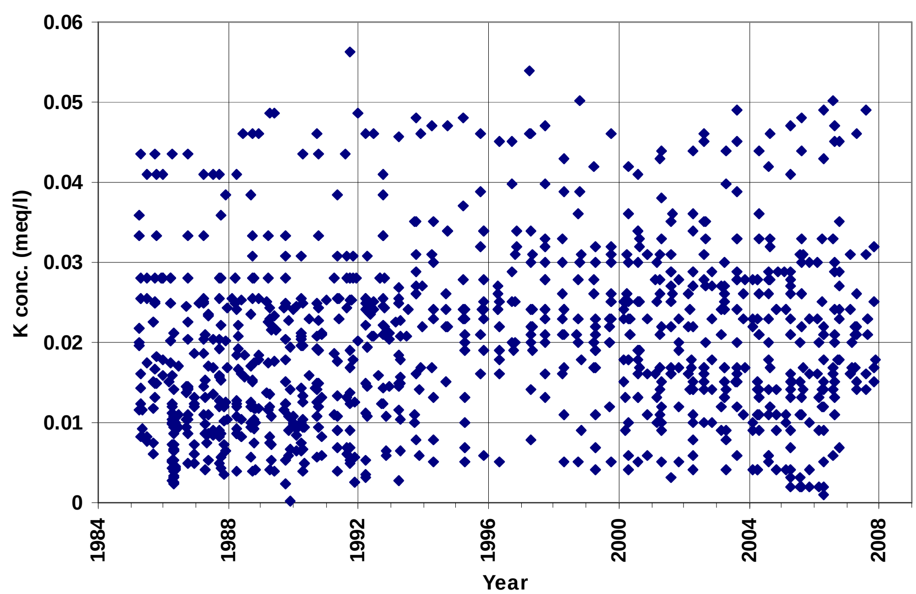


Figure 6.6 Potassium levels in groundwater sampled 1985–2007 at 19 sites in region B. Samples were normally collected on 2–6 occasions per site each year, but some longer gaps were also present in the dataset.

Because the measured potassium levels varied strongly between sampling occasions and the potential discontinuities were relatively small, our study focused on average level shifts at all 19 sites that were investigated. Figure 6.7 illustrates the annual means of the estimated trend levels when the model contained a discontinuity that was equally split between two consecutive years. One of the solid lines with attached error margins (± 2 standard errors) contained level shifts specified by the user to occur in 1990–1992, because the analytical procedure was altered in the middle of 1991. The other solid line was obtained in a purely data-driven search for the most significant discontinuity in the investigated time interval. As can be seen, the two curves differed slightly with respect to the timing of the discontinuity, whereas the size of the level shifts was practically the same in the two model runs. This shows that our change point model can also detect minor level shifts, provided that the number of sampling sites is large.

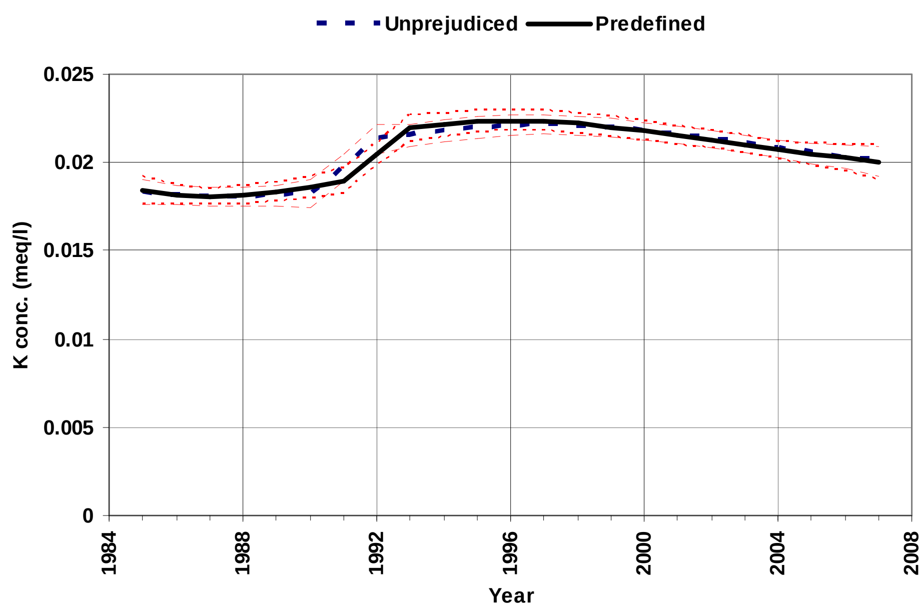


Figure 6.7 Annual means of potassium trends, including discontinuities, at 19 sites in region B. The two solid lines represent two modes of the model runs: predefined change points and unprejudiced search for discontinuities.

6.2.3 Sulphate

Figure 6.8 illustrates the results of MK tests for sulphate trends.

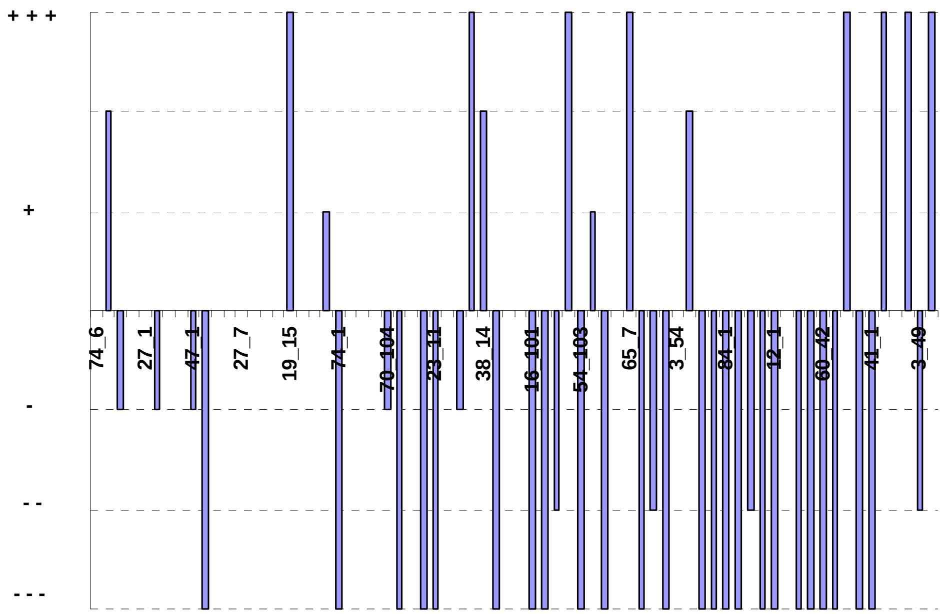


Figure 6.8 Achieved significance levels in MK tests for sulphate trends at 70 sites ordered according to median sulphate concentration. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. Three-star significances (positive and negative) were noted for (from left to right) stations 23_23, 19_15, 74_1, 58_6, 70_13, 23_11, 33_104, 16_28, 16_101, 14_15, 5_14, 54_103, 16_71, 65_7, 70_14, 16_102, 54_18, 17_10, 84_1, 13_1, 84_4, 12_1, 23_26, 69_1, 60_42, 69_10, 3_14, 21_9, 41_1, 75_2, 20_10, and 41_5.

Apparently, there were many downward trends but only a few upward trends. Closer examination of the test results revealed that there were several statistically significant downward trends in southern Sweden, particularly in hydrogeological region B (see Fig. 6.1), whereas the trends in northern Sweden were weak or nonexistent. The trends detected in region B were expected, because (i) the sulphur deposition in that part of Sweden has decreased significantly over the past decades, and (ii) shallow moraines on a primary bedrock enable rapid response to changes in deposition. Furthermore, the results of our analysis were concordant with the pronounced downward trends that were revealed when we analysed river water data from the same region.

Further examination of the sulphate levels in region B showed that the average concentration in that area decreased at about the same rate over the entire study period. However, there was substantial variation between sites, which is illustrated by the trend surface in Figure 6.9.

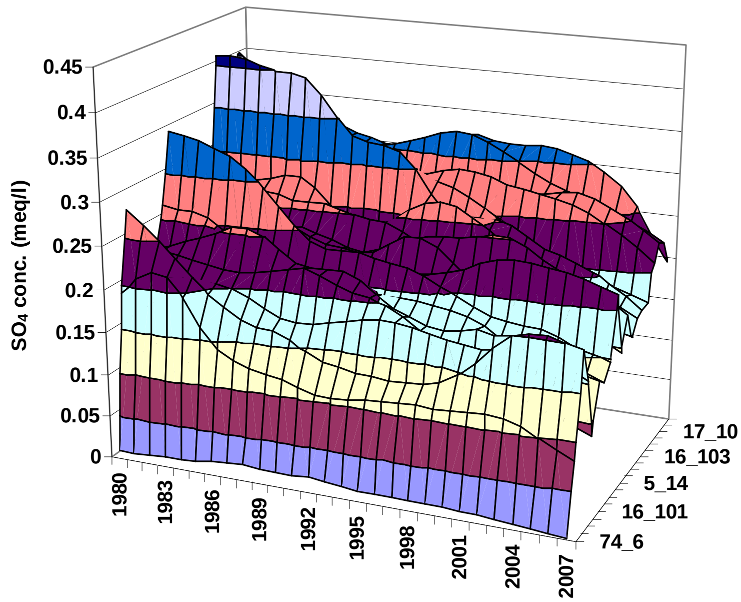


Figure 6.9 Trend surface fitted to observed sulphate concentrations at the 19 investigated stations in hydrogeological region B.

Inasmuch as repeated assessments of data quality constitute an important part of our analysis, we also searched for inexplicable level shifts in the reported sulphate concentrations. We noted that the major changes in sulphate levels seemed to be caused by natural dilution processes, because they normally coincided temporally with natural fluctuations in conductivity and other major ions. However, inspection of raw data and deviations from the fitted response surfaces also indicated a substantial serial correlation in the analysed time series. Consequently, we repeated the MK tests on data that had been reorganized in series with longer time steps. Figure 6.10 presents the results obtained when the impact of serial correlation for up to two years was suppressed. As can be seen, many significant downward trends remained.

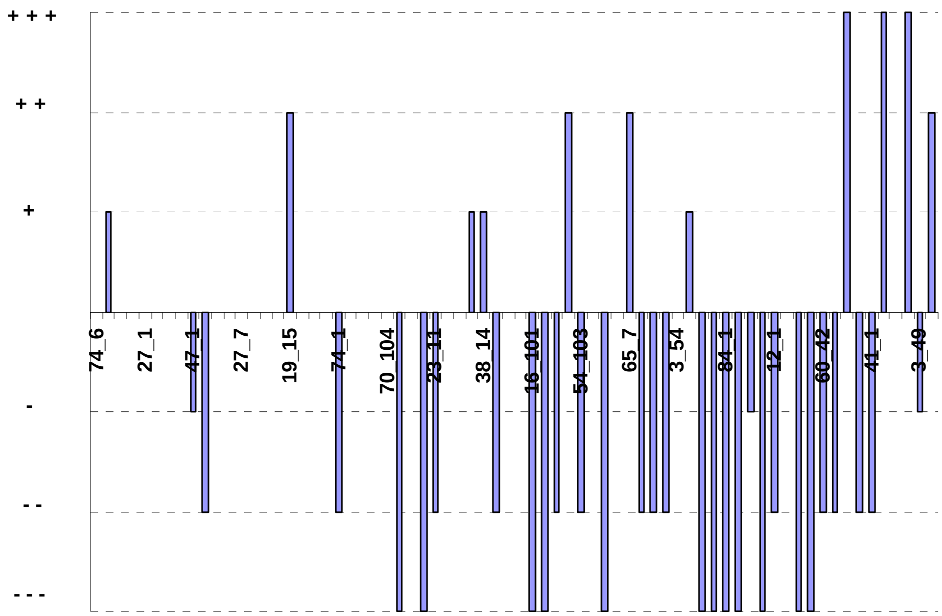


Figure 6.10 Significance in MK tests for trends in sulphate at 70 sites ordered according to median sulphate concentration, showing levels achieved when the data were reorganized into time series of biannual data. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. Three-star significances (positive and negative) were noted for (from left to right) stations 58_6, 70_13, 16_101, 14_15, 16_71, 54_18, 17_10, 84_1, 13_1, 84_4, 23_26, 69_1, 3_14, 75_2, and 20_10.

Using chloride as a covariate had approximately the same effect on the sulphate trends as on the ANC trends. Compared to the ordinary MK tests, the partial tests produced results that were almost the same, although slightly less significant, and there were considerably fewer significant trends in the ANC-to-chloride ratios.

To summarize, the sulphate data produced strong evidence of downward trends, especially in region B. However, there was no simple explanation for the spatial pattern of all downward and upward trends.

6.3 Implications for groundwater monitoring

Groundwater monitoring programmes aim to detect the impacts of human activities, which can be rather small compared to the weather-driven fluctuations and random measurement errors that influence individual observations. In addition, the relationships between causes and effects can be obscured by time lags and spurious trends. These circumstances imply two things:

(i) that assessments of regional trends must be based on relatively large networks of stations;

(ii) that the data analysis requires statistical methods for joint analysis of multiple interrelated time series.

Our assessment of Swedish groundwater data indicated that MK tests can play a key role in both exploratory data analyses and more formal tests for temporal trends. In particular, we noted that MK tests are extremely useful if they are carried out in a software package equipped with the following features:

(i) automated testing for joint trends in numerous subgroups of sampling sites;

(ii) adjustment for serial correlation;

(iii) adjustment for trends in covariates (PMK tests);

(iv) convenient handling of censored data.

Furthermore, our study demonstrated that evaluation of a relatively large and complex dataset requires efficient integration of different statistical tools. Most importantly, we found that our response surface methodology was an almost ideal complement to MK tests. Such tests proved to be efficient tools for detecting relatively small upward or downward shifts in substantial amounts of data, and our response surface methodology provided valuable information about the timing of water quality changes at different sites and the presence of artificial level shifts.

Our alkalinity and ANC study also illustrated the need for repeated assessment of data quality and the importance of performing such assessments on data from a whole network of stations. None of the data series from individual sites indicated any serious problems related to the quality of the data. As reported in Paper III, it was not until the MK tests had indicated unanticipated alkalinity trends that extensive plotting of observed data provided some clues. The final evidence of data quality problems was then provided by response surfaces and trend lines fitted to alkalinity records for low-ANC samples.

Serial correlation is another issue that must be considered in any assessment of temporal trends in environmental data. It is well known that even a moderately large autocorrelation can make the actual significance level considerably higher than the nominal level (Yue & Wang, 2004). We found that a simple generalization of the idea behind Hirsch and Slack's trend test for seasonal data is a viable alternative to the techniques currently in use. Of particular interest, our method has the advantage that it can be applied to any of the MK tests proposed in this thesis.

Chapter 5 emphasized the need for transforming the current system for monitoring surface water into an information management system. That conclusion also applies to groundwater monitoring. As national databases are merged into international databases, it will become increasingly important to remove clearly erroneous observations and to complement raw data with information about data quality in the past and the possible impact of local pollution. Because a single time series provides little information about regional trends, it would be an advantage if such an information system could also provide outputs from joint analyses of multiple time series of data.

7 Using process-based models to assess observational data

A thesis devoted to statistical evaluation of environmental monitoring data would not be complete without mentioning process-based models and their relation to observational data. Obviously, observational data are needed for calibration and validation in process-based modelling, and it is also clear that such models can be employed for spatial or temporal interpolation of observational data. However, it is not equally accepted that process-based models can contribute to assessing the quality of observational data and temporal trends therein. Neither is it widely recognized that statistical analysis of inputs and outputs from process-based models can create new roles for such models in environmental management.

This chapter briefly summarizes the relevance of process-based modelling in statistical evaluations of observational data. The INCA-N model of nitrogen flows through catchments (Wade *et al.*, 2002) is used as an example, but the conclusions are more general in nature. Parts of the discussion here are based on Paper V, and consideration is also given to complementary material derived from a research proposal coordinated by our group (Grimvall *et al.*, 2008).

7.1 The INCA-N model of nitrogen in catchments

Numerous process-oriented deterministic models have been developed to explain and predict the flow of nitrogen through catchments (e.g., Arheimer & Brandt, 1998; Refsgaard *et al.*, 1999). The INCA-N model simulates the key factors and processes that affect the amount of NO_3 and NH_4 stored in soil and groundwater systems, and it feeds the output from these systems into a multi-reach river model. The input fluxes that are taken into account in this model include the following: atmospheric deposition of ammonium and nitrate (wet and dry), application of NO_3 and NH_4 fertilizers, mineralization of organic matter (yielding NH_4), nitrification (yielding NO_3), and nitrogen fixation. From these data, various output fluxes (plant uptake, immobilization, and denitrification) are subtracted before the amount available for stream output is calculated. The final output of INCA-N consists of daily estimates of water discharge and NO_3 and NH_4 concentrations in stream water at discrete points along the main channel of the river.

7.2 Model-assisted normalization and trend assessment

Trend assessment of water quality data is a matter of separating random (weather-driven) fluctuations from more persistent changes over time. In Chapter 5, observed concentrations of total nitrogen were normalized with respect to monthly runoff by using the latter variable as a covariate in the semiparametric model described in Chapter 3. In other words, it was assumed that the average runoff during the month the sample was collected captured a substantial part of the influence of past and current weather conditions on the observed concentrations. Statistical analyses of input-output relationships in a process-based nitrogen model can suggest more relevant covariates. For example, such analyses can be used to judge the relevance of normalizing observed concentrations with respect to different combinations of contemporaneous or time-lagged runoff records.

Another approach to model-assisted normalization of observed concentrations is based on decomposing the output of a process-based model into one weather-dependent and one weather-normalized (weather-independent) component. Figure 7.1 shows how such decomposition was achieved in the study reported in Paper V. First, an observed time series of meteorological inputs to the INCA-N model was resampled to produce a collection of synthetic meteorological inputs representing the climate in the investigated catchment. More precisely, we used a form of block resampling that preserved both the seasonal pattern and the autocorrelation structure of the observed time series. Thereafter, the mean output for all synthetic inputs was regarded as weather-normalized, and the weather-specific component was isolated by subtracting this mean from the model output obtained with the actually observed meteorological forcing.

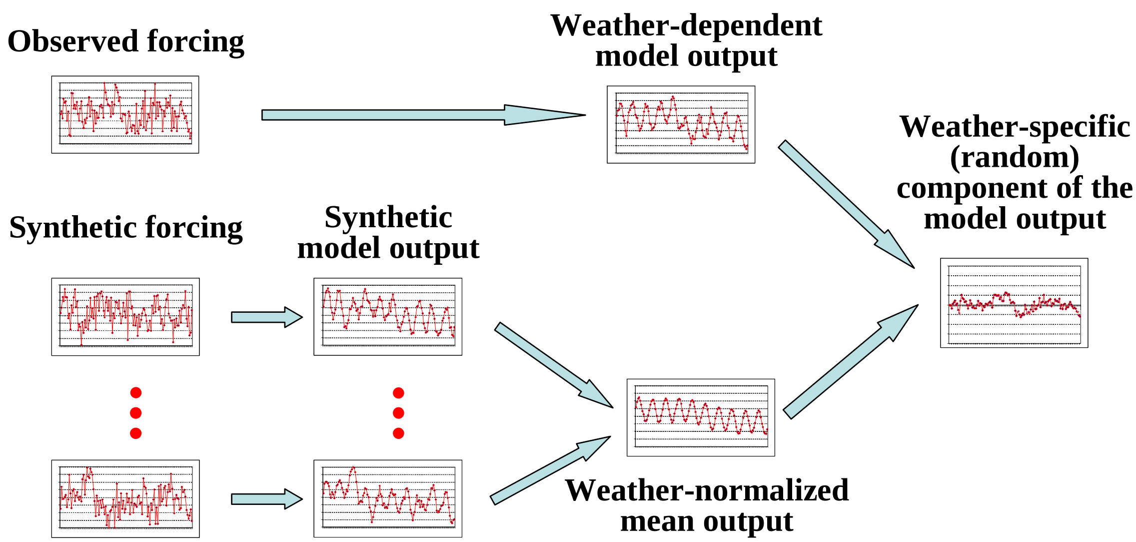


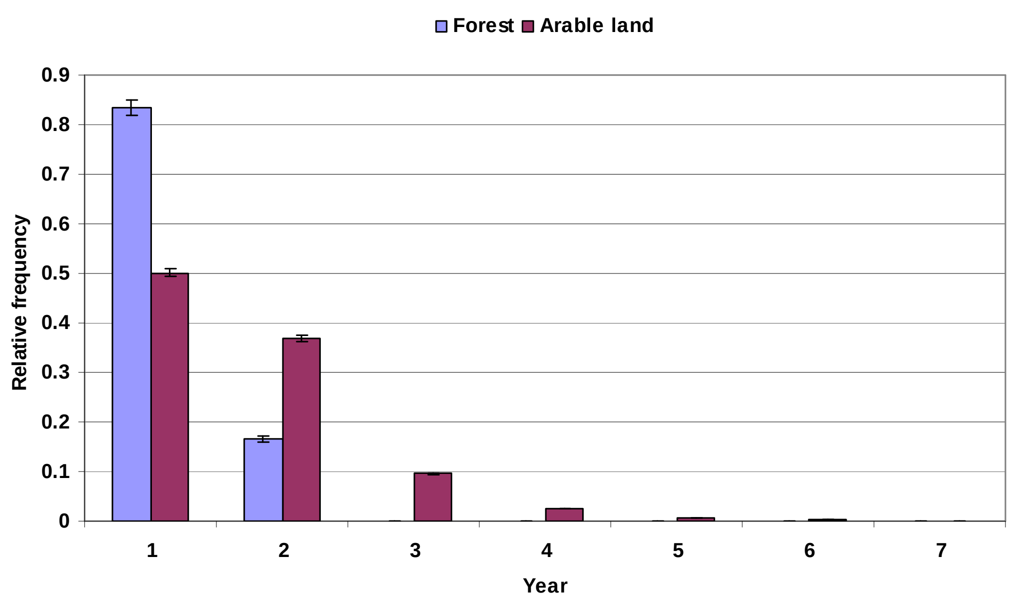
Figure 7.1 Principle of generating synthetic meteorological inputs and decomposing the output of a process-based model into a weather-specific (random) component and a weather-normalized mean function representing the effects of human interventions and normal seasonal variation.

If the process-based model were perfect, and there were no measurement errors in the chemical analysis of the water samples, we could remove the random fluctuations in an observed time series of water quality data simply by subtracting the weather-specific component derived according to Figure 7.1 from our observational data. In practice, no model is perfect. However, the weather-specific component derived from a process-based model may nonetheless function well as a covariate in the semiparametric models described in Chapters 3 and 4. Consequently, it can also help reveal trends or data quality problems that might otherwise be overlooked.

7.3 Judging the plausibility of detected trends

The statistical tests and response surface methodologies described in this thesis can reveal and quantify temporal trends and other statistical patterns in the observed data. However, identification of causes of such patterns is a matter of plausibility, which gives process-based models another important role in trend assessments. Figure 7.2 shows how our technique to generate synthetic meteorological inputs and weather-independent model outputs can clarify the possible effects of specific anthropogenic interventions, such as increased fertilizer application or atmospheric deposition, and present them in a manner that is easy to comprehend and use in a discussion regarding plausibility. In particular, such calculations can help to judge the plausibility of the magnitude and timing of temporal trends detected by statistical methods.

a)



b)

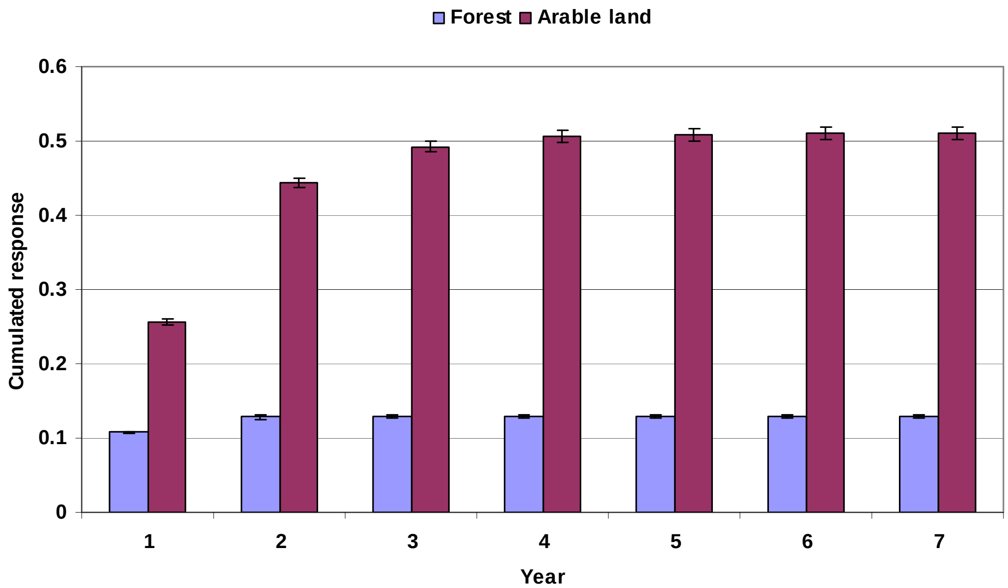


Figure 7.2 Delay in response of riverine loads of inorganic nitrogen to impulses in fertilizer inputs on arable land and atmospheric deposition on forests during the first year of the study period. The first diagram (a) illustrates the distribution of travel times for the fraction of the applied nitrogen fertilizer that is leached to water, and the second diagram (b) shows the ratio of the cumulated increase in riverine loads to the magnitude of the impulse.

The discussion of data quality problems in Chapter 5 contained a judgment of the likelihood that natural phenomena can be responsible for synchronous changes in water quality in catchments representing a very wide range of geohydrological conditions. We deemed that highly unlikely because: (i) the meteorological forcing was not synchronous in all investigated catchments, and (ii) the current process-based modelling is based on the premise that water residence time has a strong impact on water quality dynamics. If large-scale weather phenomena play a more important role than previously assumed, it should also be noted that the monitoring strategies that are currently in use need to be drastically revised. It is relatively easy to separate the impacts of human interventions from statistically independent, purely random errors in the data, whereas it is an extremely difficult task to distinguish between seemingly persistent weather effects and the influence of human activities.

7.4 Implications for water quality monitoring

The brief examples in this chapter clearly show the need for a two-way interaction between process-based modelling and the assessment of observational data. In short, not only do observational data constitute support for process-based modelling, but such modelling can also play a decisive role in the interpretation of observational data. More specifically, our investigations showed the potential of the following forms of feedback:

(i) reduction of noise in observational data;

(ii) estimation of the magnitude of human interventions;

(iii) judgment of the plausibility of synchronous increases and decreases in water quality.

8 Conclusions and final remarks

The research underlying this thesis has contributed to more efficient environmental monitoring in three respects:

(i) by providing several examples of technical improvements of statistical methods and accompanying software;

(ii) by integrating different elements of data analysis into a roadmap for retrospective analysis of multiple time series;

(iii) by drawing attention to the need for new paradigms in environmental monitoring.

8.1 Technical improvements

The technical improvements presented here aimed to meet the demand for statistical methods that can accommodate the common peculiarities of time series of environmental data and that are also easy to comprehend without being simplistic.

The MK (Mann-Kendall) tests that we refined and applied to water quality data make the family of such tests more complete. Procedures that can handle censored data were incorporated into both ordinary and partial MK tests. Furthermore, a simple method to handle serial correlations extending over more than one year was presented (Paper III), and an accompanying software package enabled automated testing of trends in various groups of user-defined data.

A versatile smoothing algorithm that can be used to search for smooth trends and synchronous increases and decreases in vector time series of data was also reported (Paper I). In particular, it can be noted that this smoothing algorithm is equipped with a new resampling procedure that can handle error terms that are correlated over time and/or across coordinates in the investigated vector time series.

A new method for the detection and estimation of abrupt level shifts in the presence of smooth trends in vector time series was also described (Paper IV). This technique unified existing approaches focusing on change points or smoothing alone.

Finally, we showed how repeated runs of process-based models can be undertaken to extract important features from temporally aggregated model outputs (Paper V). This provided an example of the need for a better integration of process-based modelling and statistical data analysis. Additional examples were briefly outlined in Chapter 7.

8.2 The importance of listening to the data

Following the device *"listen to the data,"* we integrated our methods into a roadmap for the entire pathway extending from a set of observed concentrations to conclusions about the quality of the data and existence of trends therein. Figure 8.1 shows that we made assessment of data quality a recurrent element in our analysis. The figure also illustrates how we exploited the fact that hypothesis testing and fitting of response surfaces complement each other and play different roles at different stages in the data analysis (Paper III).

At the initial stage of the analysis, the univariate MK tests and the nonparametric smoothing techniques were used as interactive exploratory tools. More specifically, we performed the following:

(i) visual inspection of *p*-values for trends in time series ordered with respect to sample means or other user-defined station characteristics;

(ii) tests for joint trends in groups of samples determined by user-defined factors or classes;

(iii) visual inspection of response surfaces in search of synchronous trends and level shifts in multiple data series.

After each step, data quality was assessed, and erroneous data were removed or corrected (Paper III).

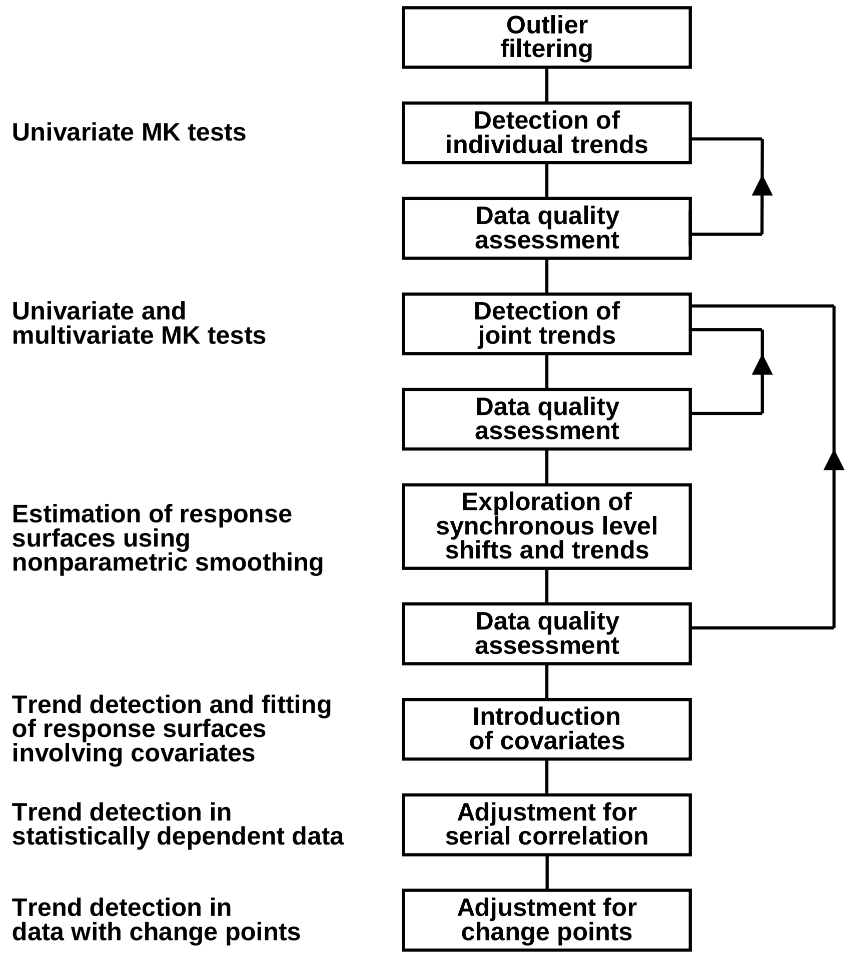


Figure 8.1 Roadmap for trend detection and assessment of data quality.

Next, we made a more formal trend analysis in which we ascertained whether the detected trends remained significant when our improved MK-tests and response surface methodologies were used to correct for covariates and correlated error terms. Finally, the presence of artificial level shifts was assessed, and, if necessary, a corrected response surface was computed.

To facilitate repeated analysis of multiple time series and to emphasize that data analysis is a much broader task than merely estimating parameters in given probability models, we also modified and extended software previously developed by our group. Practical work with the Multitest and Multitrend tools demonstrated that simultaneous analysis of multiple time series of data, and integration of visual inspection and more formal statistical analyses, constituted the key to both trend assessment and detection of flawed data.

8.3 The need for new paradigms

The datasets that were analysed in the current studies were selected to represent the best of environmental monitoring in Sweden, a country with a very long tradition of regular measurements of water quality. The records on surface water and most of the information on groundwater that we examined came from a highly reputable laboratory that has long practised state-of-the-art quality assurance. Moreover, the analytical procedures and sampling methods applied have been in use for many years, and the sampling sites have been essentially the same since the 1980s. Together, this implies that the conditions for producing excellent observational data are better in water quality monitoring than in most other fields of environmental surveillance. Nevertheless, our research revealed several remarkable problems associated with the quality of the reported data.

The level shift in potassium in groundwater (Paper IV) may be of minor practical importance. However, it is more troubling that none of the most significant temporal changes in total nitrogen (measured by persulphate digestion), total phosphorus, and TOC-to-COD ratios in surface water data from the past fifteen years could be attributed to human interventions in the environment (Papers II and IV). On the contrary, we presented strong evidence that the detected trends and level shifts were due to measurement or sampling problems. Likewise, it is less satisfactory that, according to our assessments, the strongly significant downward trends in alkalinity in acidic groundwaters were an artefact caused by poor data quality in the early 1980s (Papers III and IV).

Increased efforts to refine conventional quality assurance of reported data may reduce the problems that have been revealed, but it would be unwise to claim that such efforts can completely eliminate observational data of low quality. In addition, it is of great interest to rescue as much as possible of the information that can be extracted from existing time series of environmental quality data that obviously have artificial level shifts. This calls for a change in priorities. In Paper IV, we pointed out that climatologists have long been working on methods to detect and remove artificial level shifts in observational data (Alexandersson, 1986) and that these endeavours have recently been intensified. Similar efforts could be initiated in water quality monitoring and several other fields of environmental monitoring. The results reported in this thesis show that appropriate homogenization techniques are indeed available (Paper IV).

The dramatic advances in computer science and technology have made it feasible to make even more profound changes in the monitoring system and its interaction with the users of the collected data. First, it would be an easy task to provide more information about known or suspected data quality problems and the procedures used in the post-control of the collected data. Second, it would be possible to make available not only the originally observed data, but also various filtered datasets that have been homogenized to more correctly show the dominating overall trends.

The interaction between process-based modelling of the environment and monitoring based on data collection is another issue that deserves increased attention. Observational data are widely used to validate the former approach, but the feedback from such modelling to conventional monitoring is poorly developed. As shown in Chapter 7, both trends and artificial level shifts can be more efficiently identified if process-based models are used to estimate and remove the weather-driven fluctuations in the measured state of the environment. Finally, it should also be noted that there is substantial room for better coordination of environmental monitoring and the production of official statistics regarding the pressure on the environment. In conclusion, it is both feasible and desirable to transform the current environmental quality monitoring from a system for gathering and storing observational data to an information system that provides adequate support for environmental management.

References

Alexandersson H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology* **6**, 661-675.

Arheimer B. and Brandt M. (1998). Modelling nitrogen transport and retention in the catchments of southern Sweden. *Ambio* **27**, 471-480.

Bastviken D., Thomsen F., Svensson T., Karlsson S., Sandén P., Shaw G., Matucha M. and Öberg G. (2007). Chloride retention in forest soil by microbial uptake and by natural chlorination of organic matter. *Geochimica et Cosmochimica Acta* **71**, 3182-3192.

Butler D. (2006). 2020 Computing: Everything, everywhere. *Nature* **440**, 402-405.

Carlstein E., Do K., Hall P., Hesterberg T. and Künsch H.R. (1998). Matched-block bootstrap for dependent data. *Bernoulli* **4**, 305-328.

Caussinus H. and Mestre O. (2004). Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society Series C* **53**, 405-425.

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. *Science of the Total Environment* **258**, 149-170.

Clement L., Thas O., Ottoy J.P. and Vandrolleghem P. (2007). Data management of river water quality data: a semi-automatic procedure for data validation. *Water Resources Research* **43**.

El-Shaarawi A.H. and Niculescu S. (1992). On Kendall's tau as a test for trend in time series data. *Environmetrics* **3**, 385-411.

Foster I. (2006). 2020 Computing: A two-way street to science's future. *Nature* **440**, 419.

Grimvall A., Rahm L., Arheimer B., Carstensen J., Kauppila P. and Staskiewicz A. (2008). TEMA: Towards Evidence-based Management of coastal eutrophication. Proposal for a research project within the ERA-net Project BONUS. Theme 1: Linking Science and Policy. Linköping University, Sweden.

Hall P. and Maiti T. (2006). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics* **34**, 1733-1750.

Hamed K.H. and Rao A.R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology* **204**, 182-196.

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

Hawkins D.M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association* **68**, 941-943.

References

Helsel D.R. (2005). More than obvious: better methods for interpreting nondetect data. *Environmental Science and Technology*, October 15, 2005.

Hirsch R.M. and Slack J.R. (1984). A non-parametric trend test for seasonal data with serial dependence. *Water Resources Research* **20**, 727–732.

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.

Härdle W. (1997). Applied non-parametric regression. Cambridge University Press: Cambridge.

Lahiri S.N. (1999). Theoretical comparison of block bootstrap methods. *Annals of Statistics* **27**, 386-404.

Libiseller C. and Grimvall A. (2002). Performance of partial Mann-Kendall test for trend detection in the presence of covariates. *Environmetrics* **13**, 71-84.

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.

LiU (Linköping University) (2008). http://www.ida.liu.se/divisions/stat/research/. Accessed 2008-08-20.

Mammen E. (2000). Resampling methods for nonparametric regression. In Schimek M.G. (ed), Smoothing and Regression – Approaches, Computation, and Application. John Wiley & Sons: New York.

Miljömål (2008). http://www.miljomal.nu/english/obj3.php.
Accessed 2008-09-04.

Peterson T.C., Easterling D.R., Karl T.R., Groisman P., Nicholls N., Plummer N., Torok S., Auer I., Boehm R., Gullett D., Vincent L., Heino R., Tuomenvirta H., Mestre O., Szentimrey T., Salinger J., Førland E.J., Hansen-Bauer I., Alexandersson H., Jones P. and Parker D. (1998). Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology* **18**, 1493-1517.

Picard F., Lebarbier E., Budinska E. and Robin S. (2007). Joint segmentation of multivariate Gaussian processes using mixed linear models. Research report 5, Statistics for Systems Biology Group, Jouy-en-Josas, France.

Refsgaard J.C., Thorsen M., Jensen J.B., Kleeschulte S. and Hansen S. (1999). Large scale modelling of groundwater contamination from nitrate leaching. *Journal of Hydrology* **221**, 117-140.

Schlesinger W. (1997). Biogeochemistry. An Analysis of Global Change. Academic Press: San Diego.

SGU (Geological Survey of Sweden) (2008).
http://www.sgu.se/sgu/sv/samhalle/miljo/miljoovervakning/overvakning-grundvatten.html. Accessed 2008-08-20.

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

References

SLU (Swedish University of Agricultural Sciences) (2008). http://www.ma.slu.se. Accessed 2008-08-02.

Sonesten L. and Engblom S. (2001). Totalfosforanalyser vid Institutionen för miljöanalys 1965-2000. Institutionen för miljöanalys: Uppsala.

Srinivas V.V. and Srinivasan K. (2005). Matched block bootstrap for resampling multiseason hydrologic time series. *Hydrological Processes* **19**, 3659-3682.

Sristava M.S. and Worsley K.J. (1986). Likelihood ratio test for a change in multivariate normal mean. *Journal of the American Statistical Association* **81**, 199-204.

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

Szalay A. and Gray J. (2006). Science in an exponential world. *Nature* **440**, 413-414.

Szentimrey T. (2000). Multiple analysis of series for homogenization (MASH). Seasonal application of MASH (SAM), automatic using of meta data. In *Proceedings of the Third Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary.

Thompson M.L., Reynolds J., Cox 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.

Tukey J. (1962). The future of data analysis. *Annals of Mathematical Statistics* **33**, 1-67.

Wade A.J., Durand P., Beaujouan V., Wessel W.W., Raat K.J., Whitehead P.G., Butterfield D., Rankinen K. and Lepisto A. (2002). A nitrogen model for European catchments: INCA, new model structure and equations. *Hydrology and Earth System Sciences* **6**, 559-582.

Whitehead P.G., Wilson E.J. and Butterfield D. (1998). A semi-distributed nitrogen model for multiple source assessments in catchments (INCA): Model structure and process equations. *Science of the Total Environment* **210/211**, 547-558.

Worsley K.J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association* **74**, 365-367.

Yue S. and Wang C.Y. (2004). The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water Resources Management* **18**, 201-218.