

# **Verktyg för analys av regionala trender i grundvattnets beskaffenhet**

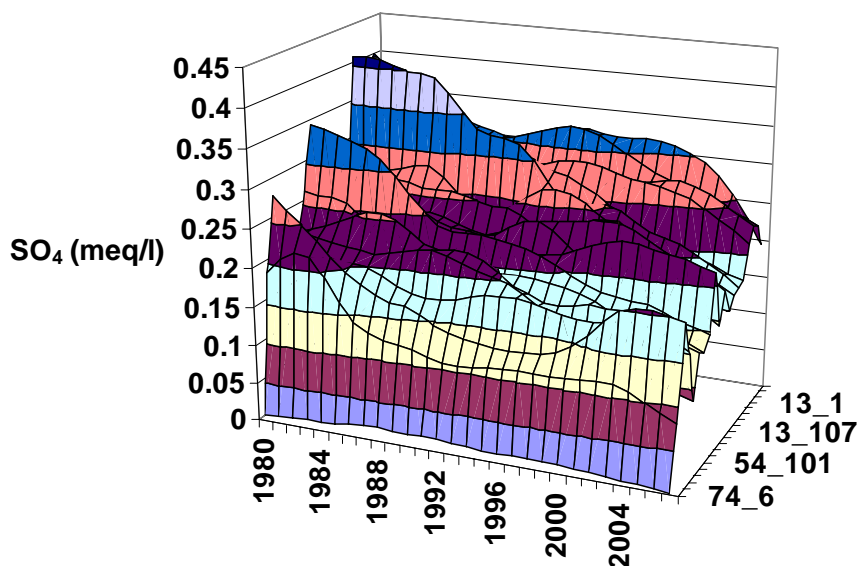
Anders Grimvall

Avd. för statistik, Institution för Datavetenskap, Linköpings Universitet

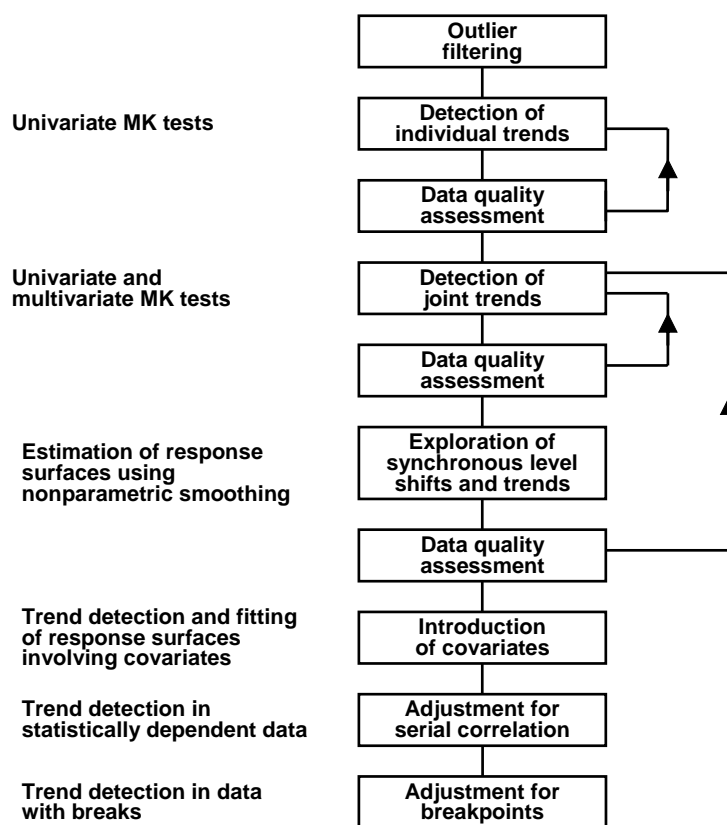
## Sammanfattning

Miljöövervakningens kanske viktigaste syfte är att upptäcka och kvantifiera långsiktiga förändringar i miljökvalitet. De statistiska metoder som idag används för detta ändamål brukar antingen vara inriktade på analyser av tidsseriedata från en mätstation i taget eller rent rumsliga analyser. Inom det aktuella projektet har vi utvecklat ny metodik och programvara för samtidig analys av tidsseriedata från flera mätstationer. Vidare har vi i samverkan med forskningsprojektet ENGO, som finansieras av Naturvårdsverket, satt samman olika statistiska metoder till en vägkarta för analys av trender och bedömning av datakvalitet. Speciellt har vi i detta projekt tagit fasta på att grundvattensdata ofta innefattar mätningar från många olika stationer och att den naturliga trögheten i många grundvattensystem gör att det krävs speciell metodik för att klarlägga om en tidstrend är statistiskt säkerställd.

Den statistiska metodutvecklingen har inriktats på responsytemetodik och icke-parametriska trendtester. Responsytorna sammanfattar de simultant skattade trendkurvorna för ett flertal stationer som ordnats efter medelkoncentrationen vid de olika stationerna (se Figur 1). De icke-parametriska testerna, som är av s.k. Mann-Kendalltyp, har i detta projekt utvidgats med en variant som kan hantera system där de naturliga svängningarna är ganska långa. Figur 2 illustrerar hur olika statistiska metoder för trendanalys och bedömning av datakvalitet kan integreras till en vägkarta för analys av grundvattenkvalitet.



**Figur 1.** Trendyta anpassad till sulfatkoncentrationer observerade vid 19 stationer inom den hydrogeologiska regionen B på sydsvenska höglandet.



**Figur 2.** Vägkarta för integrerad analys av trender och kvalitetsbedömning av tidseriedata avseende grundvattenkvalitet.

Samtliga ovanstående statistiska metoder har implementerats i VisualBasic och kan köras under Excel. Trendtesterna har samlats i ett program kallat ”*Multitest*”, medan responsymetodiken samlats i programmet ”*Multitrend*”.

När *Multitest* och *Multitrend* tillämpades på grundvattendata avseende alkalinitet, ANC (syranutraliserande kapacitet) och sulfatkoncentrationer från 1980 till 2006 konstaterades att det finns en tydlig nedåtgående trend i ANC- och sulfatdata i de regioner som tidigare haft ett betydande surt nedfall. Trenderna i alkalinitet blev missvisande eftersom mätningarna av denna vattenkvalitetsparameter hade låg kvalitet fram t.o.m. 1984.

## **Publikationer**

Huvudresultaten av projektet redovisas i följande artikel:

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

Ytterligare resultat av intresse återfinns i nedanstående två artiklar:

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

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

Dessutom har projektet bidragit till en doktorsavhandling.

Wahlin, K. Roadmap for Trend Detectyion and Assessment of Data Quality. Avd. för statistic, Inst. För Datavetenskap, Linköpings Universitet. 10 okt 2008.

Mjukvaran *Multitest* och *Multitrend* kan laddas ner från [www.ida.liu.se/stat](http://www.ida.liu.se/stat)

# Roadmap for assessing regional trends in groundwater quality

**Karl Wahlin · Anders Grimvall**

Department of Computer and Information Science,  
Linköping University, SE-58183 Linköping, Sweden

## ***Abstract***

Assessing regional trends in groundwater quality can be a difficult task. Data are often scattered in space and time, and the inertia of groundwater systems can create natural, seemingly persistent changes in concentration that are difficult to separate from anthropogenic trends. Here, we show how statistical methods and software for joint analysis of multiple time series can be integrated into a roadmap for trend analysis and critical examination of data quality. Ordinary and partial Mann-Kendall (MK) tests for monotonic trends and semiparametric smoothers for multiple time series constitute the cornerstones of our procedure. The MK tests include a simple and easily implemented method to correct for serial dependence, and the associated software is designed to enable convenient handling of numerous data series and to accommodate covariates and nondetects. The semiparametric smoothers are intended to facilitate detection of synchronous changes in a network of stations. A study of Swedish groundwater quality data revealed true upward trends in acid-neutralizing capacity (ANC) and downward trends in sulphate, but also a misleading shift in alkalinity level that would have been difficult to detect if the time series had been analysed separately.

## ***Introduction***

The awareness of large-scale and diffuse changes in the state of the environment is increasing, and this calls for efficient methods to evaluate multiple time series of data that can be more or less intercorrelated. The basic principles for analysing such data have

long been known in the statistical community (e.g., Brockwell and Davis 1996) and in several applied sciences, such as signal processing and econometrics (Griliches and Intriligator 1983; Scharf 1990). In environmetrics, analysis of joint trends in multiple time series of data was addressed more than twenty years ago (Hirsch and Slack 1984; Loftis *et al.* 1991), and there is a vast literature on methods used to model and unveil spatio-temporal patterns (Cameron and Hunter 2002; Finkenstadt *et al.* 2006; Fuentes 2002; Thompson *et al.* 2001). Nevertheless, there is substantial room for improving the procedures currently applied to evaluate environmental monitoring data collected in networks of stations. For instance, it is worth noticing that the EU guidance on ground water monitoring (Grath *et al.* 2007) does not address the fact that observations that are considered correct at the time of the sampling can be deemed erratic when more data have been collected and subjected to a thorough retrospective analysis. Here, we demonstrate how joint assessment of a large number of data series on groundwater quality can be facilitated by establishing a roadmap for regional trend analysis and providing methods and software that help coordinate exploratory analyses and formal trend testing.

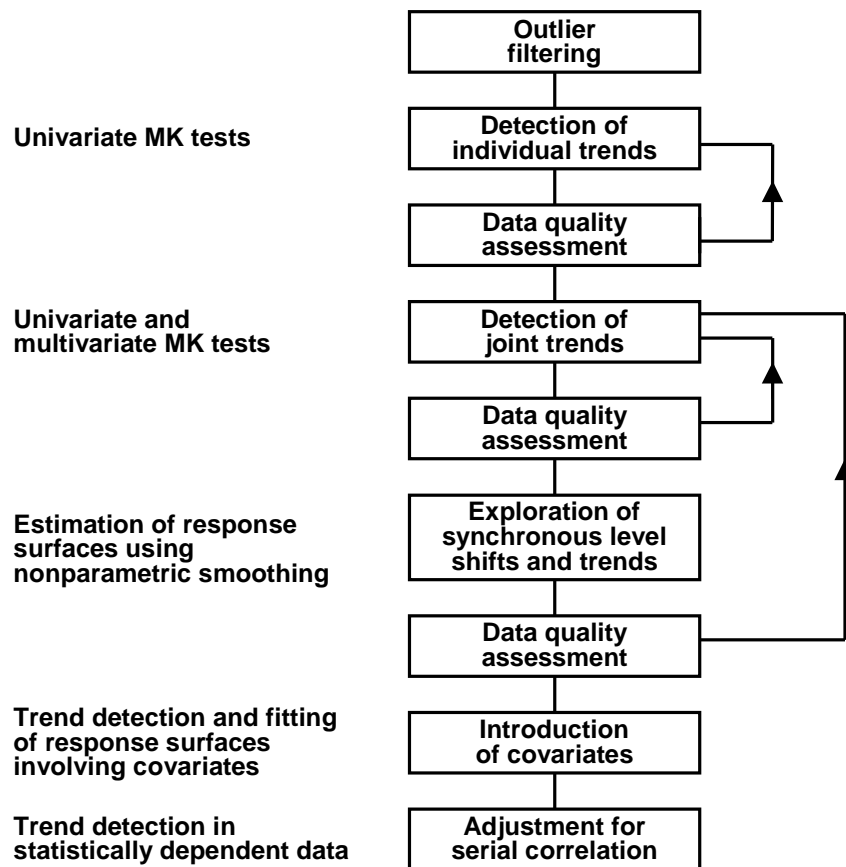
The core of the proposed roadmap for trend assessment is composed of a package of nonparametric trend tests of Mann-Kendall (MK) type and a response surface methodology that aims to explore the presence of synchronous level shifts and trends in multiple time series of data. The procedure also includes algorithms and software for multiple MK tests developed to enable automated testing for trends in user-defined groups of input data. In addition, it shows how serially correlated data and observations below the limit of quantification can be accommodated in both ordinary and partial MK tests. Response surfaces in our method are estimated using a smoothing technique that can easily be tailored to the structure of the collected data (Grimvall *et al.* 2008). In particular, we report how this technique can be applied when the data represent sampling sites that can be linearly ordered along some gradient.

To examine the performance of our strategy in assessment of regional trends, we used a dataset comprising groundwater quality data from a total of 77 stations in Sweden. This

dataset is of considerable interest in itself, because all investigated sites have been regularly sampled at least since 1980. However, it can also help determine what tools or combinations of tools that play a crucial role in the detection of regional trends and how critical assessment of data quality can be fully integrated into the statistical analysis.

### ***Roadmap for trend assessment***

Figure 1 shows that we made assessment of data quality a recurrent element in the analysis and that hypothesis testing and fitting of response surfaces are also performed repeatedly. The significance tests focus on the presence of monotonic trends. The response surfaces that are fitted to multiple series of observed data illustrate how the expected response varies over time and across sampling sites.



**Figure 1.** Roadmap for regional trend assessment.

The initial outlier filtering focuses on individual observations that differ strongly from the great majority of the other observations in the same time series. Conventional criteria, such as the number of standard deviations from the mean, can be applied to identify observations that need to be removed or corrected prior to the trend assessment. Thereafter, univariate MK tests and nonparametric smoothing techniques are used as exploratory tools. More specifically, we propose the following:

- (i) visual inspection of  $p$ -values for time series that are ordered with respect to sample means or other user-defined station characteristics (see the case study);
- (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 (Wahlin and Grimvall 2008).

After each step, data quality is assessed, and erroneous data are removed or corrected.

Next, we proceed to a more formal trend analysis in which we also take into account the impact of covariates and serial correlation. In the MK tests, covariates can be considered by adjusting the inputs prior to the tests or by performing partial trend tests (Libiseller and Grimvall 2002). In our response surface methodologies, the trend surface and the impact of covariates are estimated simultaneously (Grimvall *et al.* 2008). Finally, we ascertain whether the detected trends remain significant after corrections are made for covariates and serial correlation. In the MK tests, this can be done by reorganizing the given data into new series with longer time steps. When response surfaces are fitted to observed data, uncertainty estimates involving block resampling can reduce the impact of statistically dependent observations.

## ***Significance tests for trends***

### **Ordinary and partial MK tests**

Ordinary MK tests for monotonic trends are based on pairwise comparisons of all observations  $y_1, \dots, y_n$  in a time series, and the test statistic is given by

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

where

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

Achieved significance levels ( $p$ -values) are normally determined based on the fact that  $T$  is approximately normal with mean zero and variance  $n(n-1)(2n+5)/18$ , if  $n \geq 10$  and the null hypothesis is true, i.e., all permutations of the observed values are equally probable.

Partial MK tests are used to detect a trend in a response variable while adjusting for a trend in a covariate. If  $T$  and  $S$  denote the test statistics for trends in the response and covariate, respectively, we form the test statistic

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

where  $\hat{V}(T)$  is the estimated variance of  $T$ , and  $\hat{\rho}_{T,S}$  represents the estimated correlation of  $T$  and  $S$  (El-Shaarawi and Niculescu 1992; Libiseller and Grimvall 2002).

## **Multivariate MK tests and automated grouping of data**

The presence of a regional trend implies that sites exhibit similar, albeit not identical, trends, and this requires tests in which the evidence of increasing (or decreasing) trends is pooled for various groups of time series data. We propose significance tests based on sums of MK statistics  $T_1, \dots, T_m$  for individual time series:

$$T = T_1 + \dots + T_m$$

If the data are organized in a matrix where the rows represent years and the columns represent stations, seasons, or other groups, the null hypothesis of no trend implies that all permutations of the rows are equally probable. The columns, however, can be statistically dependent, and this can be taken into account when the variance of  $T$  is estimated (Hirsch and Slack 1984).

Because groundwater data can be grouped in many different ways, for instance with respect to sampling site, season, hydrogeological region, and other factors, it may be of interest to undertake a large number of sum tests. If the collected data can be grouped according to  $p$  factors, there is a total of  $2^p - 1$  sum tests in which univariate test statistics are summed over all levels of a subset of factors. However, some of these tests can be redundant. For example, summation over hydrogeological regions for a given station will create a redundant sum test, because each station belongs to a single hydrogeological region. Our procedure implies that all non-redundant sum tests are identified and performed.

Multivariate, partial MK tests aim to assess the presence of joint trends in several groups of data. Specifically, we assess the presence of a joint trend in the response variable that cannot be explained by a joint trend in the covariate. The test statistic will have the same form as in the univariate case, if we let  $T$  and  $S$  denote test statistics in sum tests for trends in the response and covariate, respectively. Further details about partial MK tests are given elsewhere (Libiseller and Grimvall 2002).

### **Handling of censored data**

Observations below the limit of quantification (or detection) carry information that can and should be exploited in trend tests when the measurement techniques have changed over time (Helsel 2005a). We regard all observations as intervals, i.e., pairs of real numbers. If the measured response has been quantified, the lower and upper limits of the interval coincide, or else these limits are set to zero and the limit of quantification, respectively.

If  $[a_i, b_i]$  and  $[a_j, b_j]$  are two observed intervals, representing years  $i$  and  $j$ , respectively, the sign function introduced above is modified as follows:

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

The computation of test statistics in ordinary and partial MK statistics then proceeds as usual. Analogously, the Theil slope of the trend is computed as the median of all ratios  $\frac{b_j - a_i}{j - i}$  and  $\frac{a_j - b_i}{j - i}$  for  $i < j$ . In our response surface methodology, we substitute censored observations for half the limit of quantification.

### **Adjustment for serial correlation**

Hirsch and Slack (1984) were the first to consider the impact of serial correlation on the results of MK tests. For data collected over several seasons, those investigators suggested that the raw data should be organized in a matrix in which each column represents a season, and that a sum test could be used to assess the overall trend. This idea can easily be extended to take into account serial correlation over periods longer than one year. For example, a dataset comprising observations  $y_1, \dots, y_{2n}$  made on  $2n$  consecutive years can be recoded as

Two – year period	First response	Second response
1	$y_1$	$y_2$
2	$y_3$	$y_4$
.	.	.
.	.	.
$n$	$y_{2n-1}$	$y_{2n}$

so that the statistical dependence between rows is suppressed. 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. The performance of our method to analyse data with serial correlation was examined in a simulation study (see below).

### ***Response surface methodology***

Multiple time series of data can be visualized by 3D plots in which the two horizontal axes represent time and the vector component, and the vertical axis represents the observed response (see Fig. 10). Our response surface methodology is based on the idea that, after suitable ordering of the series and an optional adjustment for covariates, the

observed responses can be approximated by a smooth function surface. The shape of the response (i.e., the temporal trend in the different vector components) is modelled in a nonparametric fashion, whereas the impact of covariates is modelled parametrically (Grimvall *et al.* 2008).

A roughness penalty approach is used along with cross-validation to adapt the degree of smoothing to the data. One smoothing parameter is employed to tune the smoothing over time, and another determines the smoothing across vector components. Explicit roughness penalty expressions have been derived for time series representing different seasons or several classes on a linear or circular scale. Here, we pay special attention to data sets representing several sampling sites that are ordered with respect to the average response at the different sites. Uncertainty bounds for the estimated response surfaces and for trend lines representing the mean response at all sites are determined by a bootstrap technique involving residual resampling. Further details about our response surface methodology have been published by our research group (Grimvall *et al.* 2008).

## **Datasets**

### **Observational data**

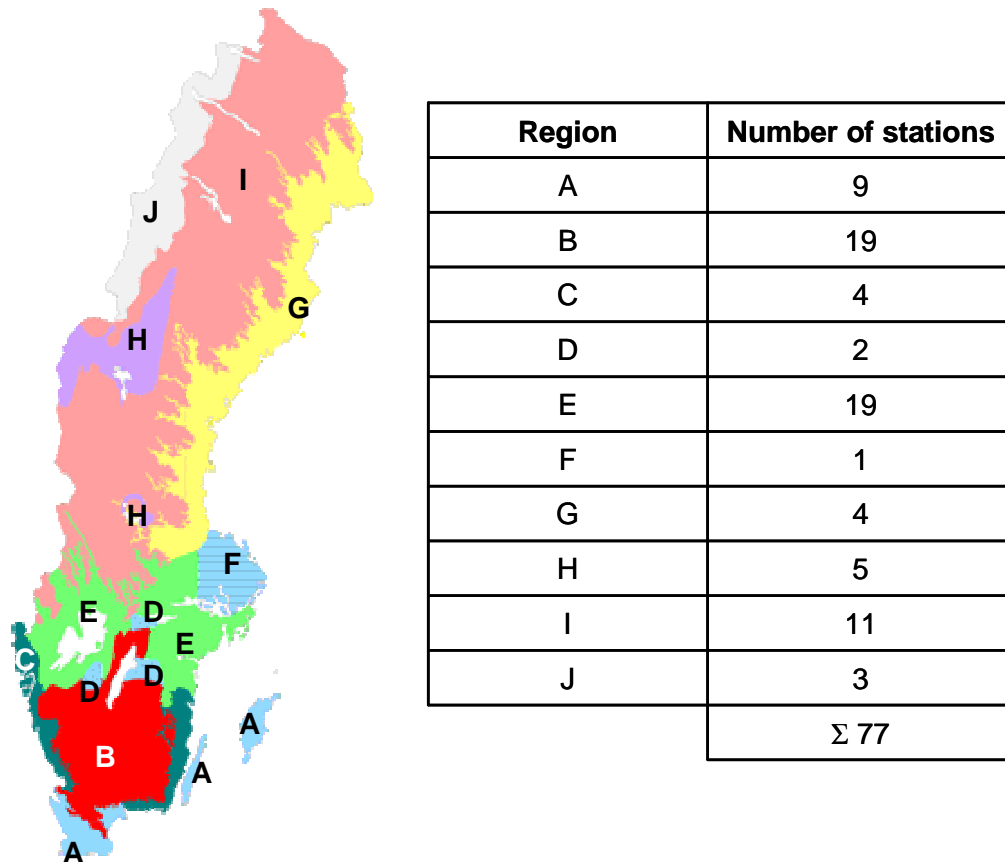
The Geological Survey of Sweden is responsible for the national monitoring of groundwater quality. Samples are normally taken 2–6 times a year, and they are subjected to analysis focused on major inorganic ions, conductivity, and temperature (SGU 2008). We investigated data from a total of 77 sites in ten hydrogeological regions (Fig. 2) where sampling has been done 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). The ANC levels were computed according to

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

Because the results raised questions about data quality (see below), we also examined sulphate, alkalinity, and ANC levels in Swedish surface waters. In those analyses, we used long time series of water quality data collected at the mouths of 37 rivers. Further

information about the national river mouth programme can be found at the website of the Swedish University of Agricultural Sciences (SLU 2008).

Finally, it should be mentioned that, since July 1992, the same laboratory has been responsible for the chemical analysis of both surface and groundwater samples collected in the national environmental monitoring programme. Before that time, the groundwater samples were analysed at two other laboratories that were commissioned from May 1980 to June 1984 and from July 1984 to June 1992, respectively.



**Figure 2.** Sweden divided into ten geographical regions based on bedrock, hydrology, and position relative to the highest coastline.

### Artificial data

Artificial groundwater quality data were generated using autoregressive (AR) models with constant or linear mean functions. The variance of the generated data was set to one,

whereas the 1-step correlation was varied from 0 to 0.4 and the slope from 0 to 0.2. The sample size was varied from 20 to 40.

## **Software**

### **MK tests**

The MK tests described above are implemented in a VisualBasic macro called Multitest (LiU 2008), which is run in Excel. Inputs are organized in tables in which the columns represent observation years, the measured responses and covariates, and factors defining region, sampling site, season, and so forth. The output of the macro comprises statistics for the following tests:

- (i) Ordinary MK tests for monotonic trends in univariate time series
- (ii) MK sum tests for joint monotonic trends in multiple time series
- (iii) Partial MK tests involving adjustment for a trend in a covariate
- (iv) Partial MK sum tests adjusting for common trends in a covariate at the investigated sites

In addition, it can be noted that the macro automatically handles censored observations and enables adjustments for serial correlation over user-defined time spans. Automatic generation of sum tests facilitates the testing for trends in groups of data or sites. The output worksheets are designed to enable simple post-processing of test results, such as sorting of  $p$ -values with respect to user-defined factors.

### **Semiparametric smoothing**

Our smoothing methodology is implemented in a VisualBasic macro denoted Multitrend (LiU 2008), which is run in Excel. Inputs are organized in tables containing one date column, one column for the response under consideration, and one or more columns for covariates. The type of smoothing (seasonal, linear or circular) is entered in UserForms. Moreover, the user can choose between different options to determine smoothing parameters and to compute uncertainty bounds by applying resampling techniques. The output of the macro comprises trend surfaces and associated uncertainty bounds. In addition, the macro computes a trend line with uncertainty bounds for the average expected response of the investigated series.

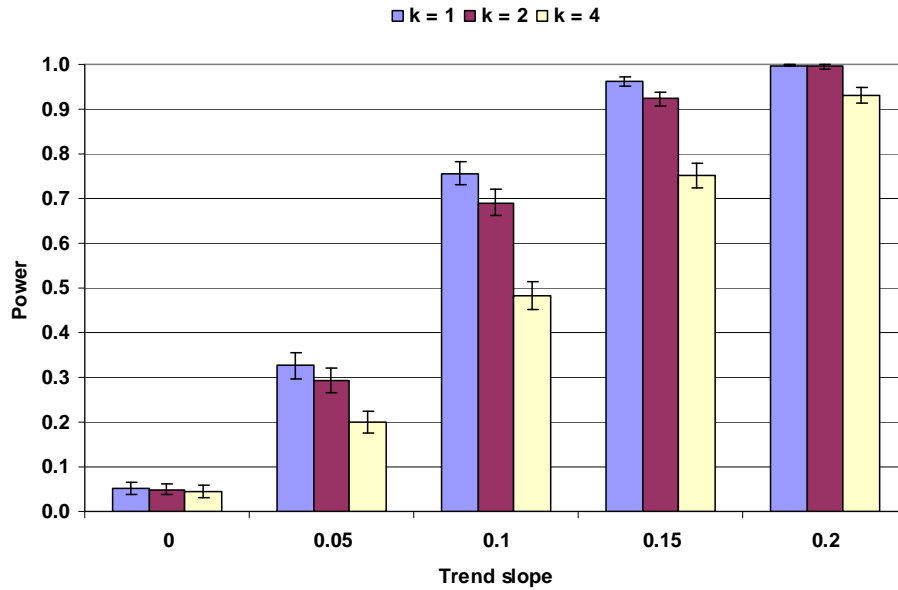
## **Results**

### **Impact of serial correlation**

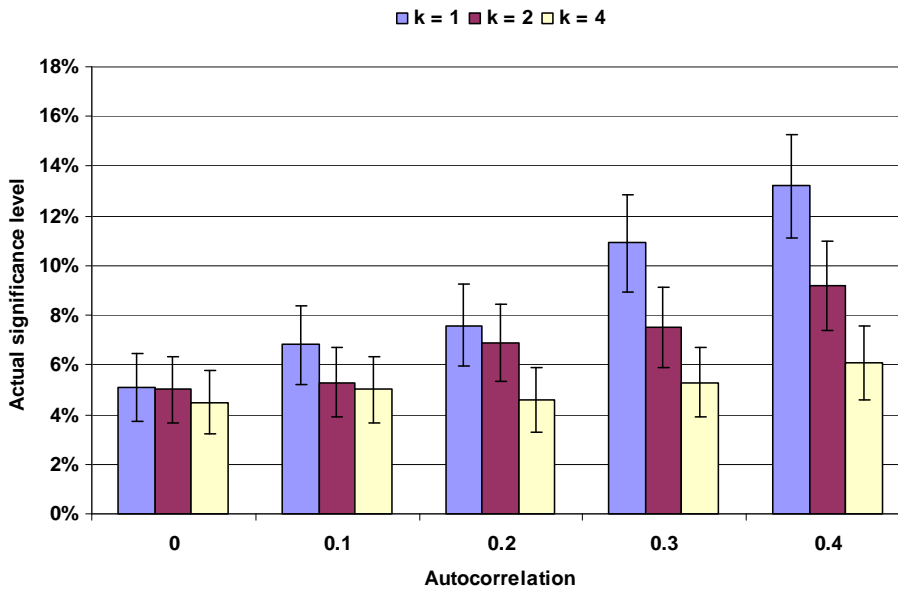
Adjustments of test statistics for serial correlation are performed to achieve better agreement between actual and nominal significance levels when the underlying data are statistically dependent. On the other hand, if data are independent, such adjustments will inevitably reduce the power of the test. As expected, our method to reorganize the given data into a larger number of shorter samples led to considerable trade-off between the accuracy of the nominal significance level and the power of the test. However, our simulations also showed that it is possible to achieve a satisfactory compromise between desirable and undesirable effects, provided the time series formed in the reorganization are at least 10 data points long.

Figure 3 shows that the loss of power was relatively small when a twenty-year time series was split into two ten-year series, each with a time step of two years, whereas the loss was more substantial when the original series was split into four five-year series, each with a time step of four years. Further simulations (not shown) demonstrated that a forty-year time series could be split into four ten-year series without substantial loss of power.

The actual and nominal significance levels of our test are identical if the autocorrelation range does not exceed the time step of the new series formed by reorganizing the original data. However, Figure 4 shows that, even if the underlying data are generated from a first order autoregressive process with a theoretically infinite autocorrelation range, our method substantially reduces the error in the nominal significance levels.



**Figure 3.** 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).

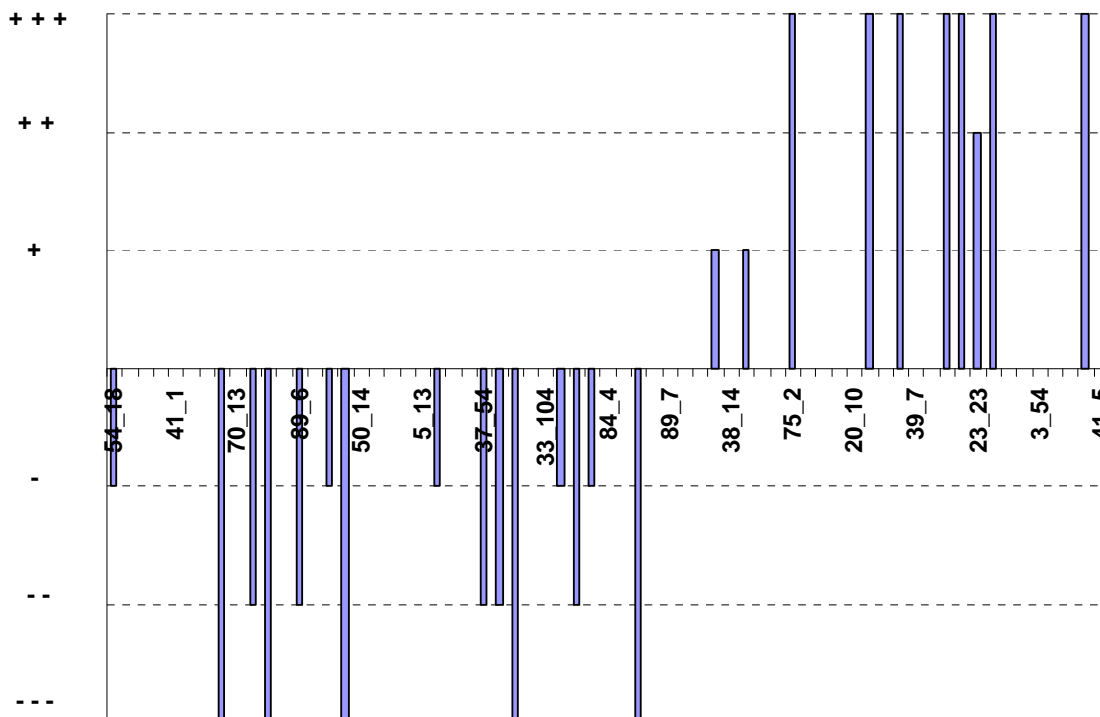


**Figure 4.** Actual significance levels of MK tests based on original and reorganized data when the original series were generated according to AR(1) processes with  $\rho = 0, 0.1, 0.2, 0.3,$  and  $0.4$ . The parameter  $k$  refers to the time step in the reorganized data series, and the nominal significance level was 5% (one-sided).

## **Alkalinity and ANC trends in groundwater**

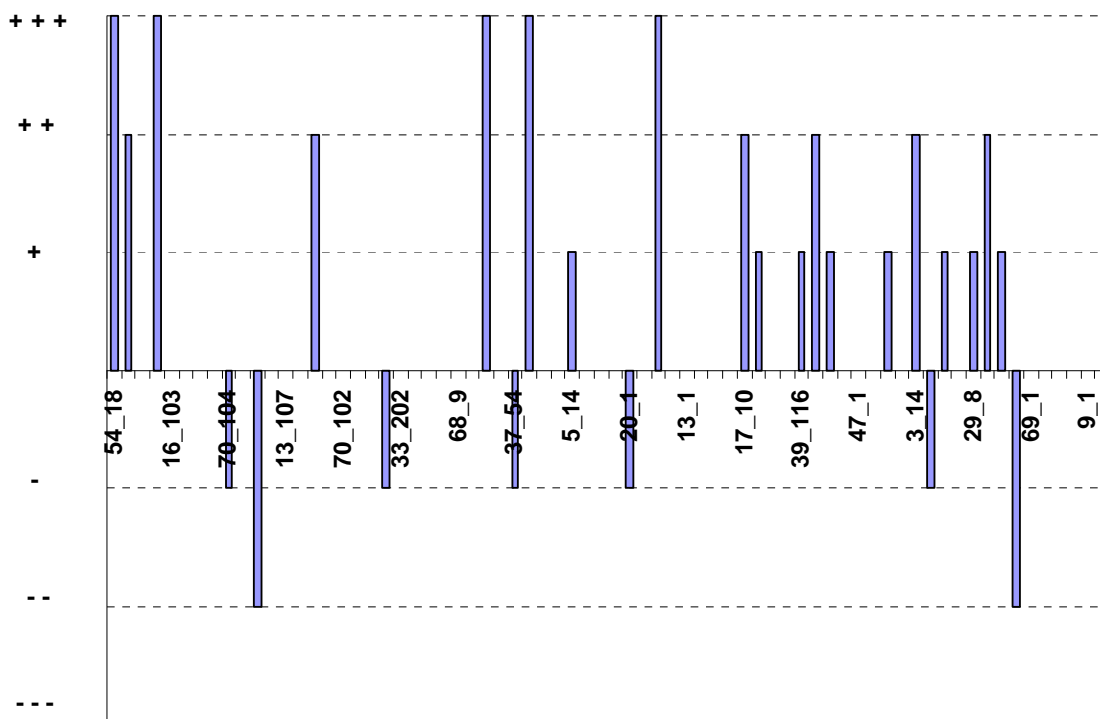
A search for outliers in the reported concentrations of major cations and anions revealed that there were obvious errors in the chemical composition of 148 of the 5,557 samples considered in the present study, and hence those samples were omitted from the trend analysis. Moreover, we excluded all data from seven of the 77 investigated sites, because both the MK statistics for temporal trends and visual inspection of collected data clearly indicated local pollution, presumably from road salt.

When ordinary univariate MK tests were again employed to examine the presence of trends in alkalinity levels, and the investigated sites were ordered according to median alkalinity, the striking pattern evident in Figure 5 emerged. As can be seen in the figure, we found significant downward trends at sites with low alkalinity and upward trends at sites with high alkalinity. The downward trends were not anticipated, because the acid deposition in Sweden has decreased considerably over the past two decades, and 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. When we performed MK tests for trends in alkalinity in 37 Swedish sampling sites, as expected, we observed the strongest upward trends in low alkalinity rivers located in regions that were previously exposed to considerable sulphur deposition.



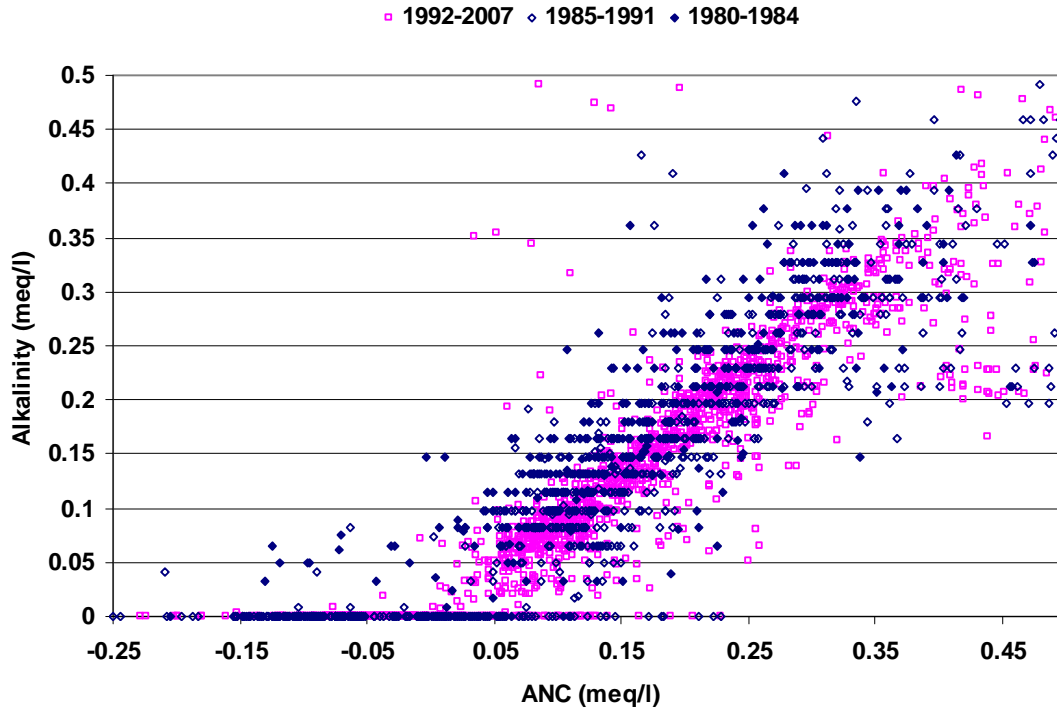
**Figure 5.** Achieved significance levels in MK tests for trends in alkalinity at 70 sites ordered according to median alkalinity. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. The station labels refer 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.

To further elucidate the existence of acidification trends in groundwater, we also examined time series of ANC levels. Figure 6 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.** 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, as seen in Figure 7, there was also a pronounced shift in the lowest alkalinity levels in 1984, when the task of analysing the groundwater samples was taken over by a new laboratory. Accordingly, 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.



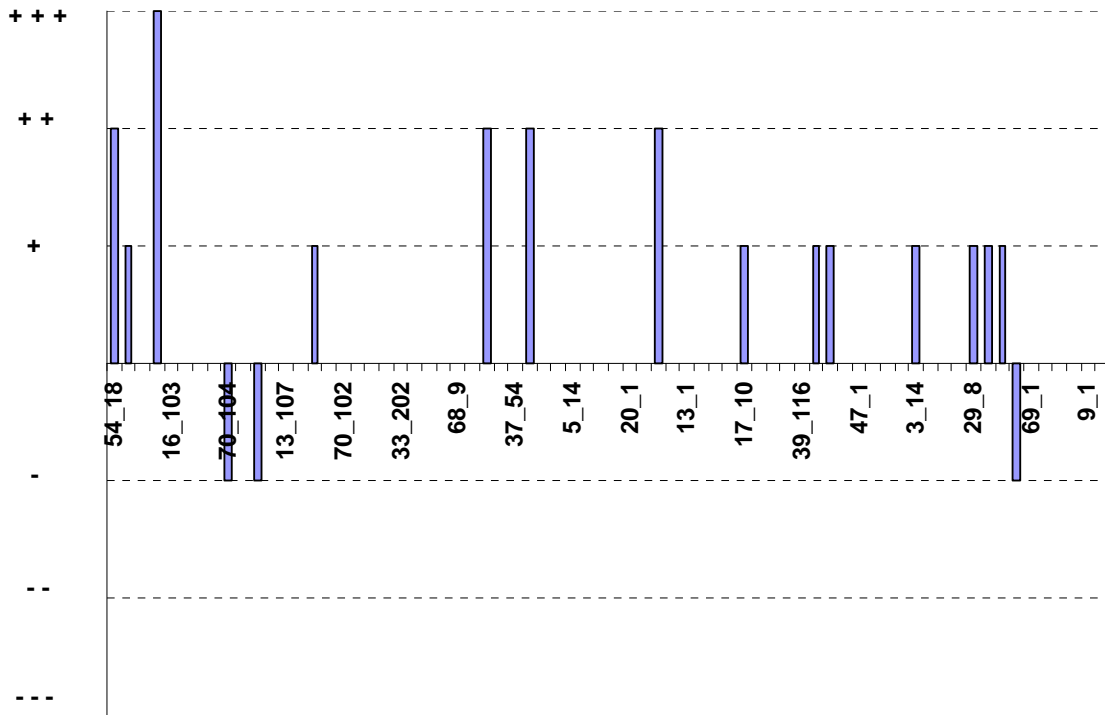
**Figure 7.** Alkalinity levels in groundwater plotted against acid neutralizing capacity. The three time periods represent data from the three different laboratories that were commissioned to perform the monitoring.

Further analysis of the ANC data revealed pronounced serial correlation for many of the investigated time series. Therefore, we also computed the achieved significance levels in MK tests where we suppressed the effect of serial correlation by reorganizing the data into biannual time series. However, as can be seen in Figure 8, 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 there were weak or nonexistent trends in northern Sweden.

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 such 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, there was considerable variation between the sampling sites.

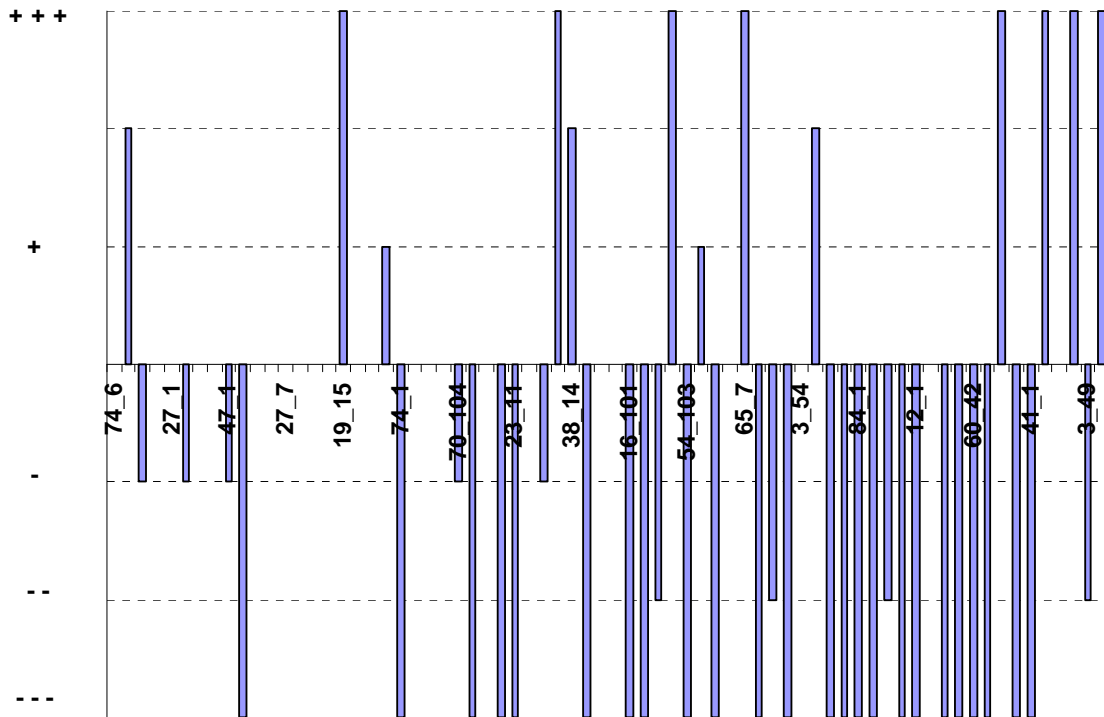


**Figure 8.** Significance in MK tests for trends in ANC at 70 sites ordered according to median ANC, showing levels achieved when the data were reorganized into time series of biannual data. Symbols: +++, ++, and + indicate positive trends significant at levels of 0.1%, 1%, and 5%, respectively; ---, --, and - signify negative trends. Three-star significance (positive) was noted for station 16\_101.

### Sulphate trends

Figure 9 illustrates the results of MK tests for sulphate trends. 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

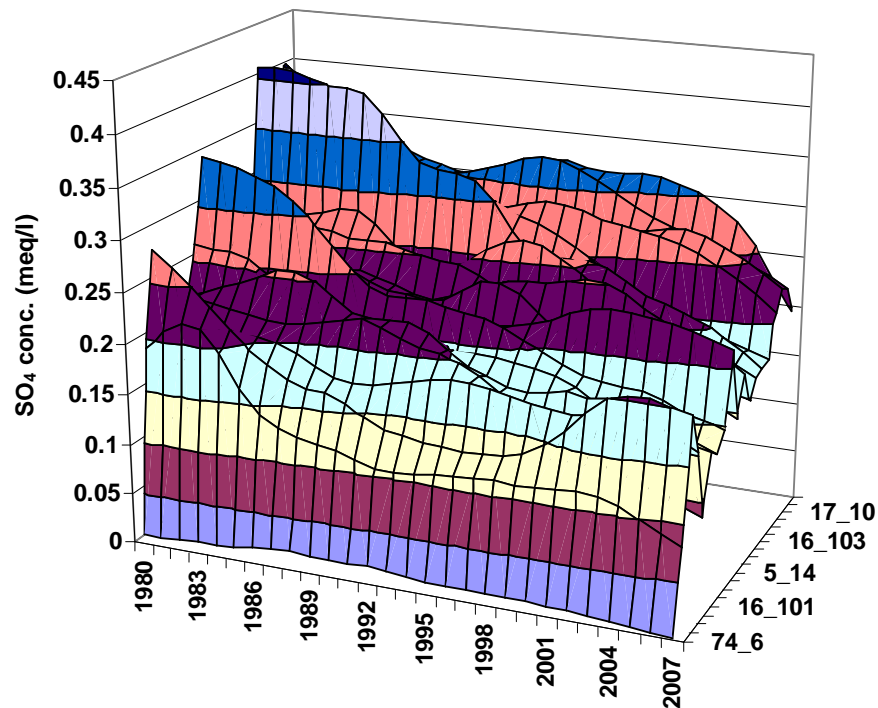
southern Sweden, particularly in hydrogeological region B (see Fig. 2), whereas the trends in Northern Sweden were weak or nonexistent. The trends detected in region B were expected, because (i) the sulphur deposition in that part of Sweden has decreased significantly over the past decades, and (ii) 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.



**Figure 9.** 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.

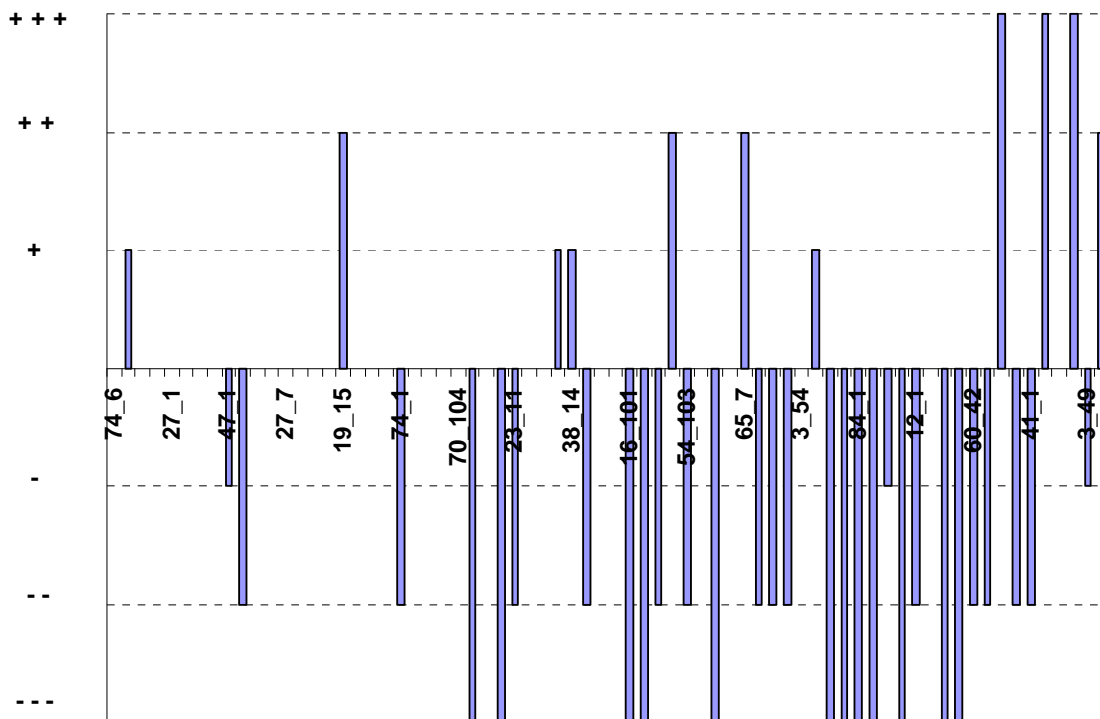
Further examination of the sulphate levels in region B showed that the average concentration in that area decreased at about the same rate over the entire study period.

However, there was substantial variation between sites, which is illustrated by the trend surface in Figure 10.



**Figure 10.** 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 roadmap, 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 11 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 11.** 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. Also, 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.

## ***Discussion and conclusions***

Groundwater monitoring programmes aim to detect human impacts that can be rather small compared to the weather-driven fluctuations and random measurement errors that influence individual observations. Accomplishment of that objective requires statistical methods that strongly suppress purely random variation, and the standard procedure is to pool data from several sampling sites and focus the statistical analysis on overall patterns in large amounts of data. We have now gone one step further by emphasizing the need for a sequence of coordinated statistical analyses that are integrated into a roadmap for simultaneous assessment of trends and data quality. The proposed collection of MK tests proved to be an efficient tool to detect 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.

Our study also showed that assessment of data quality should be repeated at all stages of the statistical data analysis. In particular, we found that examination of patterns in achieved significance levels of MK tests can effectively reveal spurious trends caused by long-lasting measurement errors. Our response surface methodology forms a natural complement to the MK tests by providing information about synchronous level shifts that may indicate changes in sampling and laboratory practices. However, it is important to note that none of the mentioned methods will separate long-lasting systematic measurement errors from actual trends. Therefore, trend analysis is also a matter of judging the plausibility of the extracted spatio-temporal patterns in the state of the environment.

The role of judgments can be illustrated with our analysis of alkalinity and ANC data. The MK tests for trends in alkalinity played a key role, because they revealed an unexpected pattern in the achieved significance levels ( $p$ -values). Furthermore, simple scatter plots showed that there was a shift in the alkalinity-to-ANC ratios of acidic samples in 1984 when a different laboratory was engaged to analyse water samples. However, we also judged that the computed ANC trends were much more plausible than the alkalinity trends. In our recent study of trends in Swedish surface waters (Wahlin and

Grimvall 2008), it was our response surface methodologies that played the most decisive role. The fitted surfaces revealed unexpectedly synchronous trends and level shifts in samples that had been taken at geographically separated sites but were analysed in the same laboratory. This observation triggered investigations that eventually led to the judgment that many time series of total nitrogen and phosphorous levels were more extensively influenced by changes in the laboratory than by actual changes in the environment. Furthermore, it is noteworthy that in both the groundwater and surface water studies the ordering of stations with respect to median concentrations helped reveal remarkable spatio-temporal patterns in the analysed data.

Standardization or normalization of environmental quality data is sometimes done to clarify temporal trends in the human impact on the environment. For example, river water quality can be normalized with respect to water discharge, and air quality with regard to various meteorological covariates (Hussian *et al.* 2004; Libiseller *et al.* 2003). Here, we compared the results obtained using ordinary MK tests and partial MK tests with chloride as covariate. In addition, we formed ANC-to-chloride and sulphate-to-chloride ratios that were subsequently analysed by ordinary MK tests. As pointed out, the use of partial tests and especially the calculation of ratios, resulted in fewer significant test results. This was expected in the present study, because (i) the peaks and troughs in ANC, sulphate, and chloride were not particularly synchronous, and (ii) the trends in chloride were generally weak at the investigated sites. In other studies, partial MK tests may provide more important information.

Serial correlation is another issue that needs to 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. A few years ago, Yue and Wang (2004) conducted a comprehensive review of the methods that have been used to adjust achieved significance levels with respect to serial correlation. In short, those authors concluded that all existing procedures have substantial shortcomings and that adjustment factors should be derived from detrended data series. 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. In particular, our method has the advantage that it can be applied to any of the MK tests proposed in the present article. Furthermore, it is not restricted to specific parametric forms of trend functions and autocorrelation functions. The performance of our method was satisfactory for autocorrelation ranges up to one tenth of the total length of the current study period.

The handling of censored data is yet another topic that needs to be addressed. We used the concepts reported by Helsel (2005a and b) and applied them to ordinary and partial MK tests, and to estimation of Theil slopes (Sen 1968; Theil 1950).

In conclusion, we have presented a set of statistical methods that address the most common problems encountered in trend analysis of groundwater quality, and we have integrated those techniques into a roadmap for such investigations. In addition, we have developed a software package that greatly facilitates joint analysis of multiple time series of data. Our case study revealed both actual trends and artificial level shifts that would have been difficult to detect if the time series had been analysed one by one.

### ***Acknowledgements***

The authors are grateful for financial support from the Geological Survey of Sweden and the Swedish Environmental Protection Agency.

### ***References***

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.

Brockwell P.J. and Davis R.A. (1996). *Introduction to time series and forecasting*. Springer: New York.

Cameron K. and Hunter P. (2002). Using spatial models and kriging techniques to optimize long-term ground-water monitoring networks: a case study. *Environmetrics*, 13, 629-656.

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.

Finkenstadt B., Held L. and Isham V. (2006). *Statistical methods for spatio-temporal systems*. Chapman & Hall/CRC: London.

Fuentes M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, 89, 197-210.

Grath J., Ward R. and Quevauviller P. (eds) (2007). *Common implementation strategy for the water framework directive. Guidance on groundwater monitoring*. Office for Official Publications of the European Communities: Luxembourg.

Griliches Z. and Intriligator M.D. (eds) (1983). *Handbook of econometrics*. Elsevier: Amsterdam. <http://www.sciencedirect.com/science/handbooks/15734412>. Accessed 28 June 2008.

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

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

Helsel D.R. (2005b). Insider censoring: distortion of data with nondetects. *Human and Ecological Risk Assessment*, 11, 1127-1137.

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.

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., Grimvall A., Waldén J. and Saari H. (2003). Meteorological normalisation and non-parametric smoothing for quality assessment and trend analysis of tropospheric ozone data. *Environmental Monitoring and Assessment*, 100, 33-52.

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

Loftis J.C., Taylor C.H. and Chapman P.L. (1991). Multivariate tests for trend in water quality. *Water Resources Bulletin*, 24, 505-512.

Scharf L. (1990). *Statistical signal processing*. Prentice Hall: New Jersey.

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

Sen P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379-1389.

SGU (Geological Survey of Sweden) (2008).

<http://www.sgu.se/sgu/sv/samhalle/miljo/miljoovervakning/overvakning-grundvatten.html>. Accessed 2008-08-20.

SLU (Swedish University of Agricultural Sciences) (2008).

<http://www.ma.slu.se>. Accessed 2008-08-20.

Theil H. (1950). A rank-invariant method of linear and polynomial regression analysis, I, II, and III. *Nederlandsche Akad. van Wetenschappen Proc.*, 58, 386-392, 521-525 and 1397-1412.

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.

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

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.

# Estimating artificial level shifts in the presence of smooth trends

K. Wahlin,<sup>1</sup> A. Grimvall,<sup>1</sup> S. Sirisack<sup>2</sup>

<sup>1</sup>Department of Computer and Information Science, Linköping University, SE-58183  
Linköping, Sweden

<sup>2</sup>Department of Mathematics, National University of Laos, Vientiane, Laos  
e-mail: [angri@ida.liu.se](mailto:angri@ida.liu.se)

## ***Abstract***

Changes in observational data over time can be severely distorted by errors in measurements, sampling, or reporting. Here, we show how smooth trends in vector time series can be separated from one or two abrupt level shifts that occur simultaneously in all coordinates. Trends are modelled nonparametrically, whereas abrupt changes and the impact of covariates are modelled parametrically. The model is estimated using a back-fitting algorithm in which estimation of smooth trends is alternated with estimation of regression coefficients for covariates and assessment of sudden level shifts. The proposed method is adaptive in the sense that the degree of smoothing over time and across coordinates is controlled by a roughness penalty and cross-validation procedure that automatically identifies the interdependence of the analysed data. Furthermore, it uses a resampling technique that can accommodate correlated error terms in the assessment of the uncertainty of both smooth trends and discontinuities. The method is applied to water quality data from Swedish national monitoring programmes to illustrate how known discontinuities can be quantified and how previously unrecognized discontinuities can be detected.

## ***Introduction***

The concern about global warming and other long-term changes in the environment has led to increased interest in long time series of environmental data, and there is also a

growing awareness of issues related to data quality. A variety of guidance documents for environmental monitoring have been prepared, and substantial efforts have been made to assess and assure the quality of the reported data (e.g., Aguilar *et al.* 2003; Grath *et al.* 2007). Even so, trends in observational data can be severely contaminated by errors in measurements, sampling, or reporting. Compilations of regional and global climate datasets have revealed numerous examples of abrupt changes caused by things like new instrumentation or relocation of sampling sites (e.g., Jones 1995; Klein Tank *et al.* 2002). Our own investigations of air and water quality have indicated that, in such data, artificial level shifts can be a substantial problem even if state-of-the-art quality assurance is applied (Libiseller *et al.* 2005; Wahlin and Grimvall 2008a). Accordingly, there is a strong need for statistical methods that can help detect and estimate discontinuities and other inhomogeneities, and rescue information from old monitoring data.

So far, the most systematic attempts to assess the homogeneity of time series of environmental data have been undertaken by climate scientists. The statistical methods used in that field have their roots in a likelihood ratio test for shift in level at some unknown instant (Hawkins 1977; Worsley 1979), as well as a multivariate extension of that test (Sristava and Worsley 1986). In his pioneering work, Alexandersson (1986) embedded the mentioned type of tests in a procedure in which the climate signal of a candidate series is first removed by subtracting a reference series that is known to be homogeneous. More recent methods aim to detect inhomogeneities without specifying a priori that some series are more reliable than others. Szentimrey (1997) developed algorithms in which each series is compared with an optimally weighted mean of the other series, and Caussinus and Mestre (2004) designed a decision algorithm based on pairwise comparisons of data series from neighbouring sites. Picard and co-workers (2007) showed more generally how numerical algorithms for maximum likelihood estimators in mixed linear models can be employed to simultaneously estimate an arbitrary number of change points in a data matrix.

Outside the climate sector, efforts to detect artificial level shifts in environmental data have been less systematic. We have recently emphasized that joint analysis of multiple

time series is needed to efficiently detect and estimate inhomogeneities (Wahlin and Grimvall 2008b). However, the tools developed by climatologists are far from ideal for estimating abrupt level shifts in vector time series in which the coordinates have different trends. Therefore, the aim of the current study was to reduce that deficiency by developing methods for joint analysis of smooth trends and synchronous discontinuities in multiple data time series. In this article, we first describe our model for detection and estimation of inhomogeneities. Thereafter, we show how the model parameters can be estimated, and, finally, we apply our technique to surface and groundwater quality data from Swedish national monitoring programmes.

## ***Models and algorithms***

### **A model class for level shifts in the presence of smooth trends**

Let us consider an  $m$ -dimensional vector time series

$$\mathbf{y}_t = (y_t^{(1)}, \dots, y_t^{(m)})^T, \quad t = 1, \dots, n$$

representing observations made at  $m$  sites at  $n$  equidistant time points. Our model can then be written

$$\mathbf{y}_t = \boldsymbol{\alpha}_t + \sum_{k=1}^p (\mathbf{x}_{kt} - \bar{\mathbf{x}}_k) \boldsymbol{\beta}_k + \boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, n$$

where  $\boldsymbol{\alpha}_t$ ,  $t = 1, \dots, n$ , is a deterministic trend surface,  $\mathbf{x}_{kt}$ ,  $t = 1, \dots, n$ ,  $k = 1, \dots, p$ , a set of  $p$  vector time series of covariates,  $\boldsymbol{\gamma}_t$ ,  $t = 1, \dots, n$ , a sequence of vectors that are stepwise constant in each coordinate, and  $\boldsymbol{\varepsilon}_t$ ,  $t = 1, \dots, n$ , a sequence of random vectors with mean zero and constant covariance matrix. As is customary, the symbols  $\bar{\mathbf{x}}_k$ ,  $k = 1, \dots, p$  represent sample means, and

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T) = (\boldsymbol{\beta}_1^{(1)}, \dots, \boldsymbol{\beta}_1^{(m)}, \dots, \boldsymbol{\beta}_p^{(1)}, \dots, \boldsymbol{\beta}_p^{(m)})^T$$

denotes a time-independent vector of regression coefficients.

The sequence  $\gamma_t$ ,  $t = 1, \dots, n$ , can be parameterized in different manners in different applications. If observed data have a single change point common to all coordinates, we can set

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

where  $\theta(j)$  denotes the level shift in the  $j$ th coordinate between time  $t_1$  and  $t_1+1$ . The parameter  $\mu$ , which is unidentifiable in the presence of the vectors  $\alpha_t$ ,  $t = 1, \dots, n$ , is normally selected so that the sum of all  $\gamma_t^{(j)}$  is zero. Furthermore, we can introduce simple parametric forms of the level shifts, such as

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

or

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

Functions with two or more change points are defined analogously. If it is suspected that the measured data have been biased during a certain time period, it may be of interest to use the following parameterization:

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

Expressions of the form

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

where  $0 < \delta < 1$  can be useful if a level shift takes place in two consecutive steps.

### **Point estimation of parameters and selection of smoothing factors**

Because the models presented above are over-parameterized, it is necessary to introduce some constraints or regularization functions when the parameters are estimated.

Following the ideas previously outlined by our group (Grimvall *et al.* 2008), we used a penalized least squares technique in which we minimized an expression of the form

$$S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{t=1}^n \sum_{j=1}^m (y_t^{(j)} - \hat{y}_t^{(j)})^2 + \lambda_1 L_1(W_1, \boldsymbol{\alpha}) + \lambda_2 L_2(W_2, \boldsymbol{\alpha})$$

where

$$L_1(W_1, \boldsymbol{\alpha}) = \sum_{(t_1, j_1, t_2, j_2, t_3, j_3) \in W_1} \left( \alpha_{t_1}^{(j_1)} - \frac{\alpha_{t_2}^{(j_2)} + \alpha_{t_3}^{(j_3)}}{2} \right)^2$$

$$L_2(W_2, \boldsymbol{\alpha}) = \sum_{(t_1, j_1, t_2, j_2, t_3, j_3) \in W_2} \left( \alpha_{t_1}^{(j_1)} - \frac{\alpha_{t_2}^{(j_2)} + \alpha_{t_3}^{(j_3)}}{2} \right)^2$$

and  $\hat{y}_t^{(j)}$  denotes a prediction of  $y_t^{(j)}$  based on data for all time points  $s \neq t$ . The vector  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$  contains two nonnegative factors controlling the smoothness of the trend surface, and  $W_1$  and  $W_2$  denote two different smoothing patterns. Normally,  $W_1$  is set to

$$W_1 = \{(t, j, t-1, j, t+1, j), \quad t = 2, \dots, n-1 \quad j = 1, \dots, m \}$$

in order to generate horizontal smoothing (smoothing over time), whereas  $W_2$  is used to impose a vertical smoothing pattern (smoothing across coordinates). However, the model can accommodate any user-defined smoothing patterns  $W_1$  and  $W_2$  (Grimvall *et al.* 2008). If both  $\lambda_1$  and  $\lambda_2$  are large, the fitted smooth trend surface will be almost a plane in  $\mathbf{R}^3$ . If both  $\lambda_1$  and  $\lambda_2$  are small, the smoothing of observed values will be practically negligible.

We achieved global minimization of  $S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$  by systematically searching for  $\boldsymbol{\lambda}$ -values that made the prediction error sum of squares (*press*) as small as possible when we left out one year of observations at a time, estimated the model parameters using the remaining data, and summed the squared prediction errors for the observations that were left out. Furthermore, it can be noted that the minimization of  $S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$  with respect to  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\gamma}$  for given smoothing factors was accomplished by employing a back-fitting algorithm alternating between estimation of  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\gamma}$  respectively. Written as

pseudocode, our algorithm to determine *press* for given  $\lambda$ -values had the following structure:

---

**Back-fitting algorithm for joint estimation of smooth trends and discontinuities**

---

1. Initialize  $\boldsymbol{\gamma} \equiv 0$
2. Initialize  $\boldsymbol{\beta}$  by using a multiple linear regression model with intercept to regress  $\mathbf{y}$  on  $\mathbf{x}_1, \dots, \mathbf{x}_p$
3. Initialize *press* = 0
4. Initialize  $s = 1$
5. Repeat
6.  $T = \{1, \dots, s-1, s+1, \dots, n\}$   
Cycle

$$u_t^{(j)} = y_t^{(j)} - \sum_{k=1}^p \beta_k^{(j)} (x_{kt}^{(j)} - \bar{x}_k^{(j)}) - \gamma_t^{(j)}, \quad t \in T, \quad j = 1, \dots, m$$

$$\arg \min_{\boldsymbol{\alpha}} \left[ \sum_{t \in T} \sum_{j=1}^m (u_t^{(j)} - \alpha_t^{(j)})^2 + \lambda_1 L_1(W_1, \boldsymbol{\alpha}) + \lambda_2 L_2(W_2, \boldsymbol{\alpha}) \right]$$

$$u_t^{(j)} = y_t^{(j)} - \alpha_t^{(j)} - \gamma_t^{(j)}, \quad t \in T, \quad j = 1, \dots, m$$

$$\arg \min_{\boldsymbol{\beta}} \left[ \sum_{t \in T} \sum_{j=1}^m (u_t^{(j)} - \beta_0 - \sum_{k=1}^p \beta_k^{(j)} (x_{kt}^{(j)} - \bar{x}_k^{(j)}))^2 \right]$$

$$u_t^{(j)} = y_t^{(j)} - \alpha_t^{(j)} - \sum_{k=1}^p \beta_k^{(j)} (x_{kt}^{(j)} - \bar{x}_k^{(j)}), \quad t \in T, \quad j = 1, \dots, m$$

$$\arg \min_{\boldsymbol{\gamma}} \left[ \sum_{t \in T} \sum_{j=1}^m (u_t^{(j)} - \gamma_t^{(j)})^2 \right]$$

$$\gamma_t^{(j)} \leftarrow \gamma_t^{(j)} - \text{mean} \left( \sum_{t \in T} \sum_{j=1}^m \gamma_t^{(j)} \right)$$

until the relative change in the penalized sum of squares on  $T$  is below a pre-specified threshold

$$\textit{press} \leftarrow \textit{press} + \sum_{j=1}^m (y_s^{(j)} - \alpha_s^{(j)} - \sum_{k=1}^p \beta_k^{(j)} (x_{ks}^{(j)} - \bar{x}_k^{(j)}) - \gamma_s^{(j)})^2$$

$$s \leftarrow s+1$$

7. until  $s = n$
- 

If the number of observations varies from cell to cell in the matrix defined by time and site, it is easy to modify the formulae indicated above. Further details are given elsewhere (Grimvall *et al.* 2008).

## Uncertainty assessment

The uncertainty of the fitted trend surface and the estimated level shifts was assessed using a residual resampling technique introduced by Grimvall *et al.* (2008). As in ordinary residual resampling in regression models with non-random design, the covariates  $\mathbf{x}_{kt}$ ,  $t = 1, \dots, n$ ,  $k = 1, \dots, p$ , were kept fixed, and new response values were generated by setting

$$y_t^{*(j)} = y_t^{(j)} - e_t^{(j)} + e_t^{*(j)}, \quad t = 1, \dots, n \quad j = 1, \dots, m$$

where  $e_t^{*(j)}$  denotes a resampled residual, and the same symbol without an asterisk denotes the original residual (Mammen 2000). However, after selecting new residuals by sampling with replacement, pairs of resampled residuals were swapped until the correlation pattern was similar to that of the original residuals. Further details are available elsewhere (Grimvall *et al.* 2008).

## Computational aspects

The algorithm presented above has been implemented as a VisualBasic macro for Excel (LiU 2008). Extensive experiments in which the macro was tested on simulated data with known level shifts showed that the back-fitting invariably converged to solutions that were coherent with the true data model. In addition, our runs with real water quality data did not reveal any convergence problems. The computational burden varied strongly with the mode in which the algorithm was run. Fitting a model with given smoothing factors and without resampling for uncertainty assessments took less than a second for the datasets presented in this article, and this was achieved mainly by exploiting the band matrix structure of the system of  $mn$  linear equations providing estimates of the  $\alpha$ -parameters (Hussian *et al.* 2004; Stålnacke and Grimvall 2001). Moreover, the cross-validation was non-problematic, because the exact levels of the selected smoothing factors did not have a significant impact on the results obtained. The computational effort in the resampling is more substantial. When we generated 200 sample replicates and allowed 100,000 residual swaps for each replicate, the computational time varied from less than a minute to almost an hour when our datasets were analysed on a standard PC.

However, if the error components are only weakly correlated, the number of swaps, and hence also the total computational time, can be substantially reduced.

## ***Case studies of change-point detection***

### **Observational data**

We tested the methods and algorithms on surface and groundwater data from national monitoring programmes conducted in Sweden. The surface water data represented one site in Lake Vänern (Dagskärsgrundet) and samples collected close to the mouths of fifteen major rivers in the northern part of the country (Table 1). The statistical analysis focused on total phosphorus, TOC (total organic carbon), and COD (chemical oxygen demand) measured as permanganate consumption. Datasets and further information can be obtained from the Swedish University of Agricultural Sciences (SLU 2008).

**Table 1.** The investigated rivers and their recipients: the Bothnian Sea (BS) and the Bothnian Bay (BB)

<b>River</b>	<b>Recipient</b>	<b>River</b>	<b>Recipient</b>
Torne	BB	Ångermanälven	BS
Kalix	BB	Indalsälven	BS
Råne	BB	Ljungan	BS
Lule	BB	Delångersån	BS
Pite	BB	Ljusnan	BS
Ume	BS	Gavleån	BS
Öre	BS	Dalälven	BS
Gide	BS		

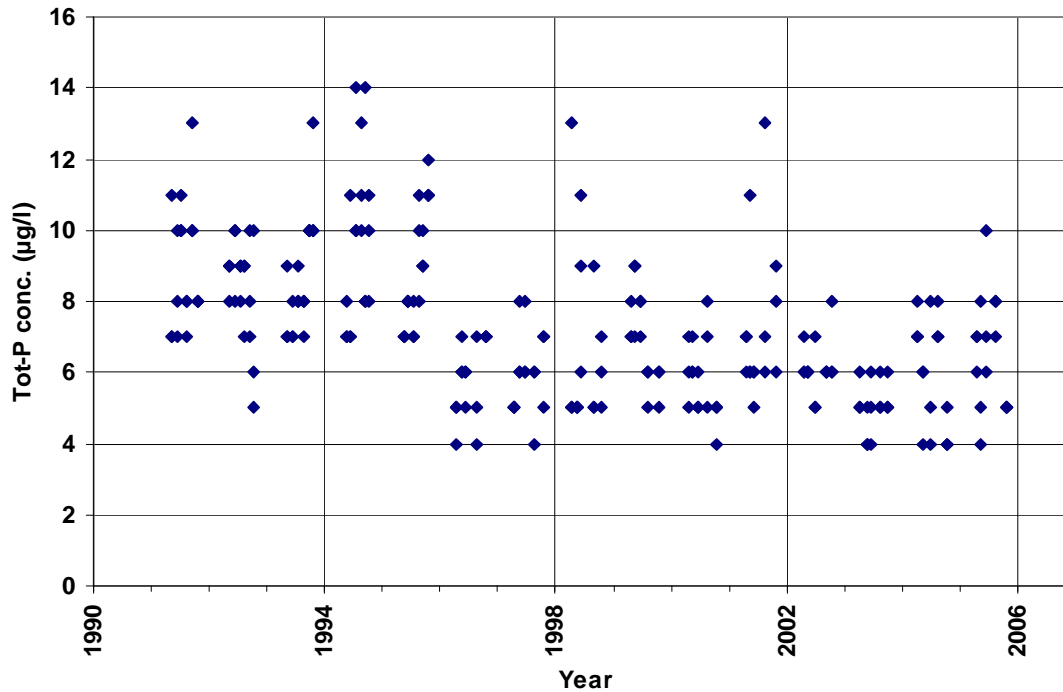
The groundwater represented a total of 77 sites. Special attention was paid to reported levels of potassium and alkalinity, and records of acid neutralizing capacity (ANC) computed according to

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

Datasets and further information about the monitoring programme can be obtained from the Geological Survey of Sweden (SGU 2008).

## Level shifts at known instants

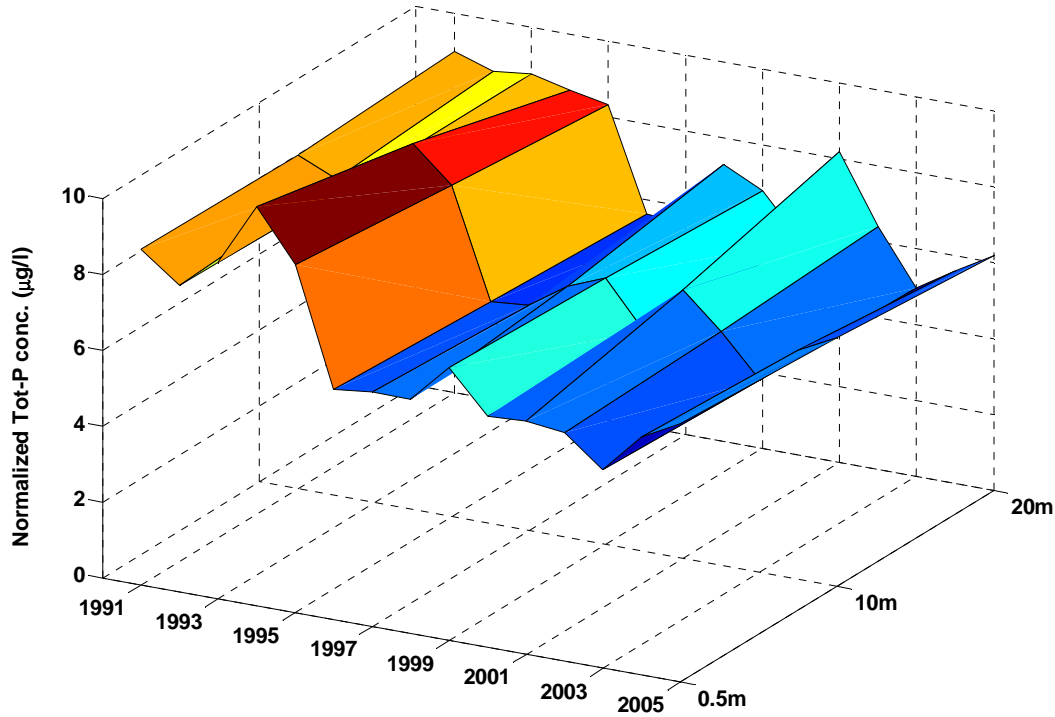
Visual inspection of Figure 1 indicates that a level shift in the reported phosphorus concentrations took place in 1996, after the procedure to correct for the blank level of the chemical analysis was altered. As expected, it was impossible to achieve a good fit to this dataset when our model was run with large smoothing factors and without any discontinuities.



**Figure 1.** Total phosphorus (Tot-P) levels in surface water at Dagskärsgrund in Lake Vänern, 1991–2005. Samples were collected on 4–6 occasions per year from April to October at depths of 0.5, 10, and 20 m.

This was also the case when water temperature was incorporated as a covariate, and Figure 2 illustrates the rather rough trend surface that was selected by our algorithm. We subsequently augmented our model with a discontinuity between 1995 and 1996, and the level shift was assumed to be of the same size at all sampling depths. This modification substantially improved the fit to reported data. Moreover, the cross-validation then indicated that the highest predictivity of the model was obtained for large values of the smoothing factors (Fig. 3). The size of the discontinuity was estimated to 3.1 µg/l, and

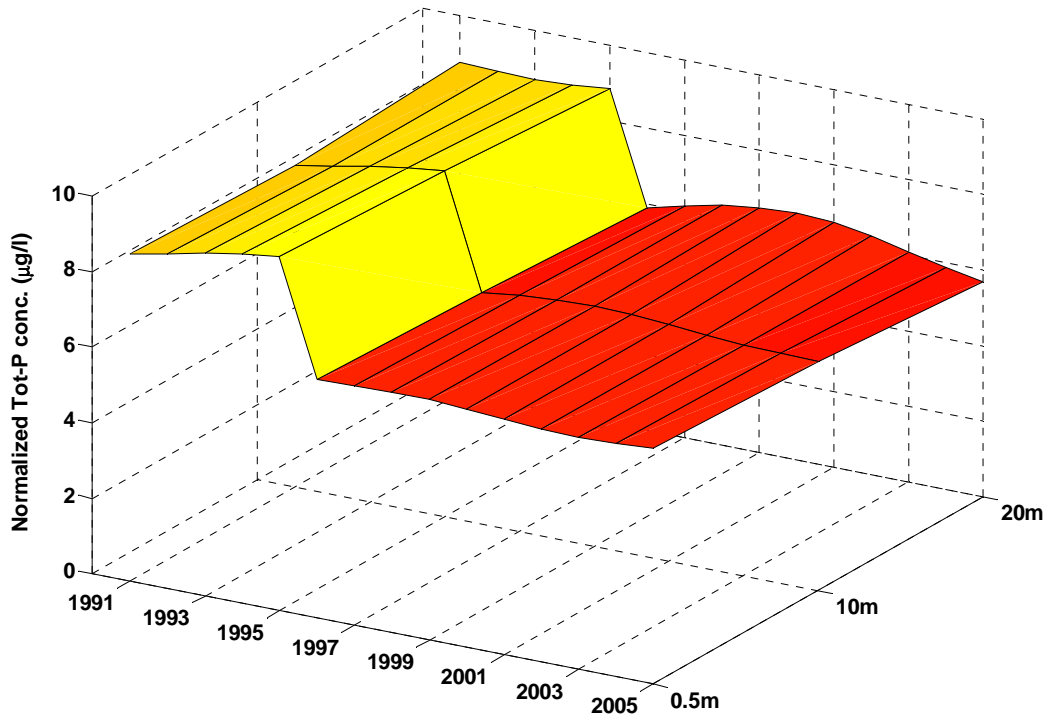
residual resampling showed that the standard error of the estimated level shift was considerably smaller ( $0.44 \mu\text{g/l}$ ).



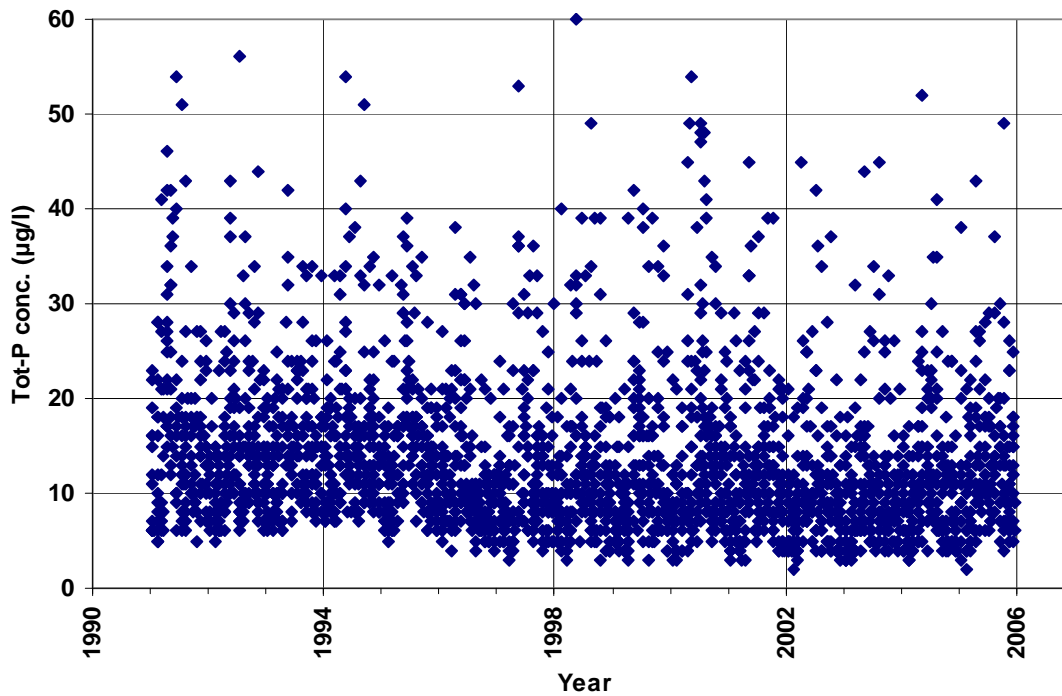
**Figure 2.** Trend surface without discontinuities for the total phosphorus (Tot-P) levels shown in Figure 1. Cross-validation indicated that the optimal smoothing factors were  $\lambda_1 = 0.16$  and  $\lambda_2 = 32$ .

In a recent study (Wahlin and Grimvall 2008a), we found that there were also abrupt level shifts in other phosphorus records from the same laboratory. Figure 4 shows the measured concentrations of phosphorus in fifteen major rivers in northern Sweden. When we reanalysed that dataset using the algorithm presented here, we found that the level shift in 1996 was statistically significant, and that discontinuity emerged even more clearly when the analysis was restricted to the four rivers with the lowest frequency of outliers. Figure 5 shows the fitted trend surface with the estimated discontinuity. Inasmuch as the change in the 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. Table 2 illustrates the estimated level shifts and their standard errors. In particular, it can be noted that level shifts also occurred in rivers where measured phosphorus concentrations were far above the detection limit of the analytical procedure employed.



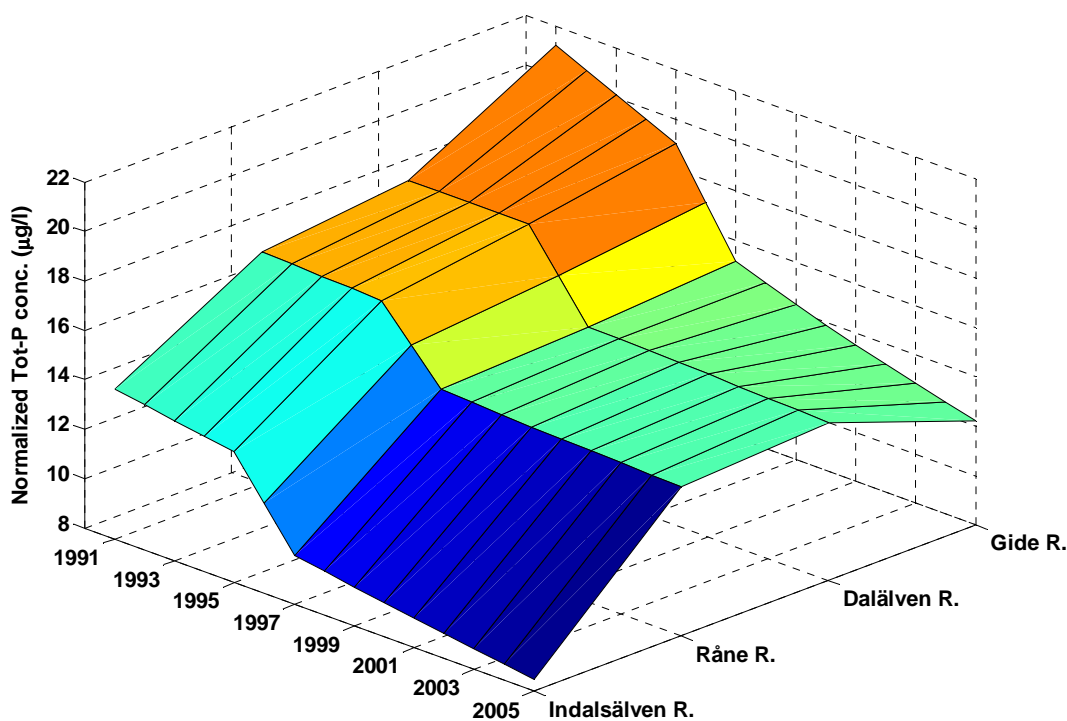
**Figure 3.** Smooth trend surface augmented with a discontinuity between 1995 and 1996. The underlying data were the same as in Figures 1 and 2, and cross-validation indicated that the optimal smoothing factors were  $\lambda_1 = 10240$  and  $\lambda_2 = 16$ .



**Figure 4.** Total phosphorus (Tot-P) levels recorded at the mouths of fifteen major rivers in northern Sweden. Monthly sampling was done in all rivers throughout the investigated period.

**Table 2.** Estimated level shifts in total phosphorus data from four rivers in northern Sweden. The model had level shifts that were equally split between 1995–96 and 1996–97, and the size of the 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



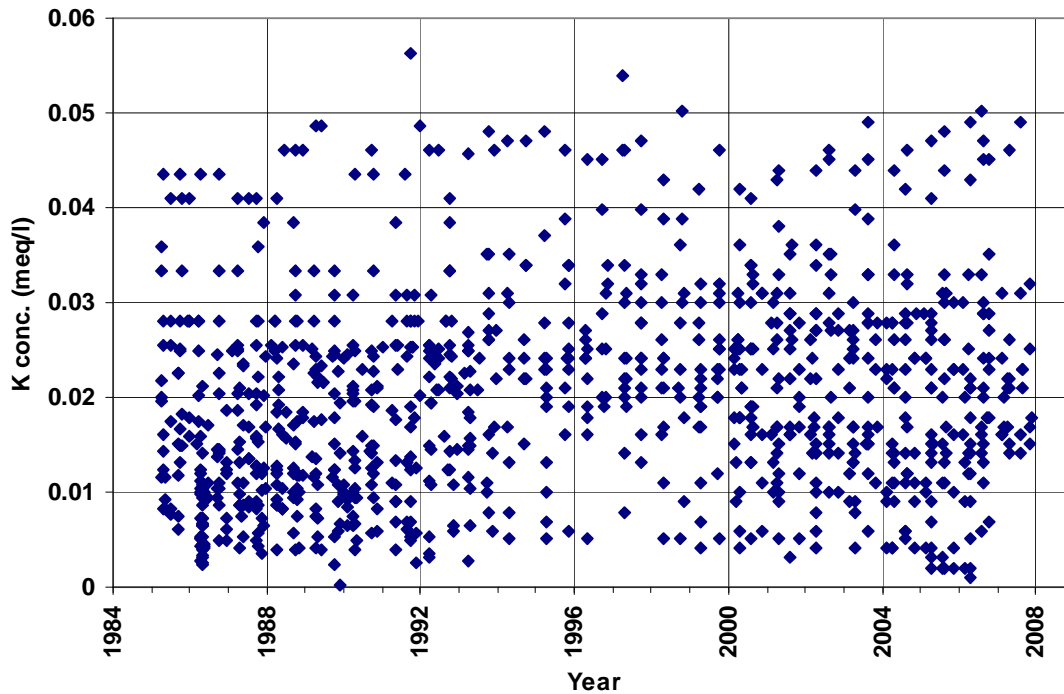
**Figure 5.** Trend surface with discontinuities fitted to total phosphorus (Tot-P) concentrations in four major rivers in northern Sweden. The statistical model and the sampled rivers were the same as in Table 2.

### Level shifts at unknown instants

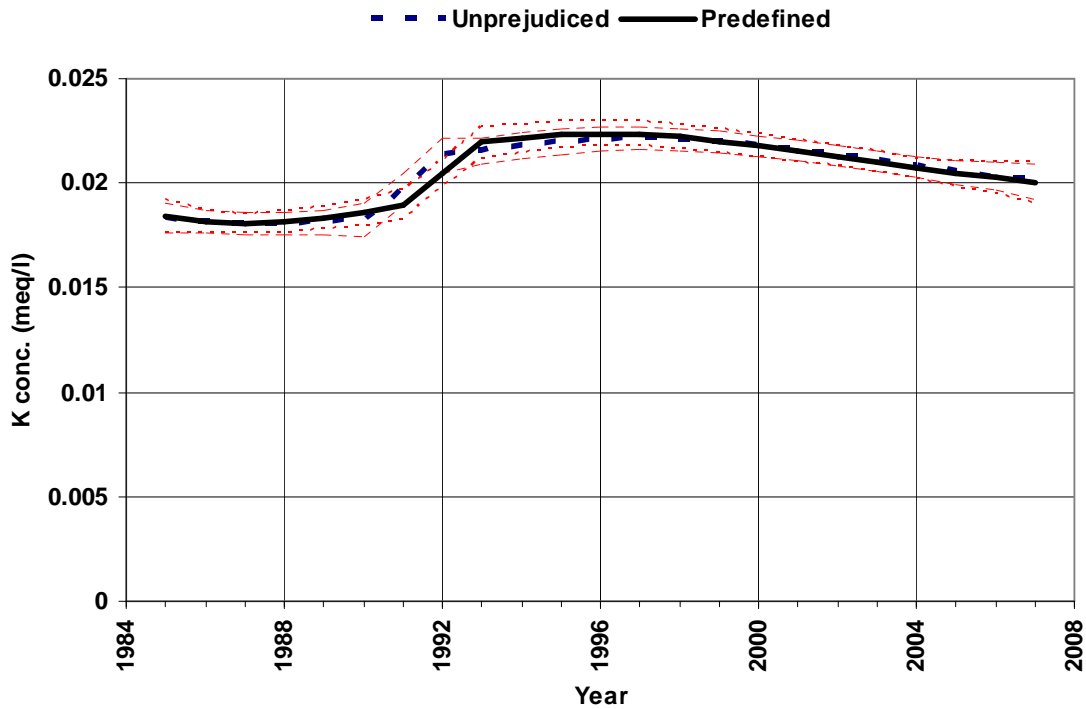
The response surfaces in Figures 3 and 5 were obtained with change points specified by the user. However, the results were identical when the algorithm was run with an unprejudiced search for level shifts, and that outcome was expected because the abrupt changes were quite evident. Figure 6 illustrates a dataset in which the presence and location of discontinuities is less obvious.

Since the measured potassium levels varied strongly between sampling occasions, and the potential discontinuities were relatively small, we focused our study on average level shifts. Figure 7 illustrates the annual means of the estimated trend levels when the model contained a discontinuity that was equally split between two consecutive years. The thick

solid line with attached error margins ( $\pm 2$  standard errors) contained two consecutive level shifts specified by the user to occur in 1990–1992, because the analytical procedure was altered in the middle of 1991. The thick dashed line was obtained in a purely data-driven search for the most significant discontinuity in the investigated time interval. As can be seen, these two curves differ 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 was expected, considering that the timing can be strongly influenced by a relatively small number of observations that are temporally close to the true change point. The size of the level shift is less sensitive to small subsets of observations, provided that the smoothing factors are not too small.

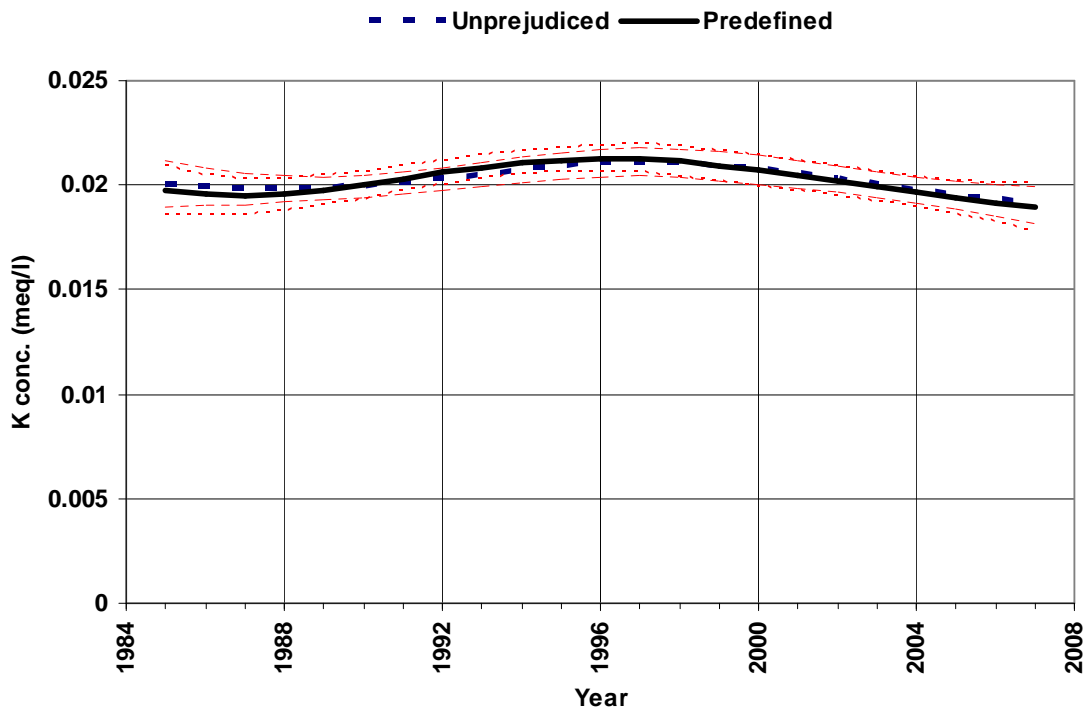


**Figure 6.** Potassium concentrations in groundwater sampled in 1985–2007 at 19 sites in the South Swedish Highlands. Samples were normally collected on 2–6 occasions per year at each site, although there were also some longer breaks in the dataset.



**Figure 7.** Annual means of potassium trends, including discontinuities, at 19 sites in the South Swedish Highlands. The thick solid and the thick dashed line represent two modes of the model runs: predefined change points and unprejudiced search for discontinuities, respectively.

Figure 8 shows the trend lines obtained after the estimated level shifts were removed. Apparently, there were only minor differences between the results obtained with a user-defined change point and those acquired in an unprejudiced search for discontinuities.



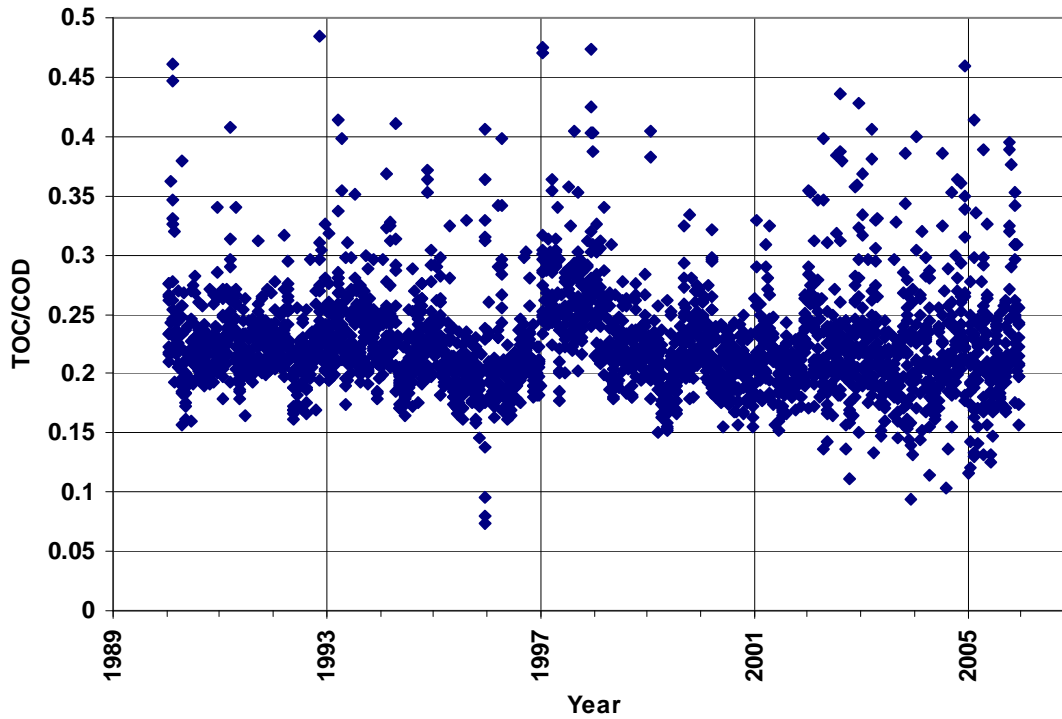
**Figure 8.** Annual means of potassium trends after removing the estimated level shifts. The thick solid and the thick dashed line represent two modes of the model runs: predefined change points and unprejudiced search for discontinuities, respectively.

### Temporary bias

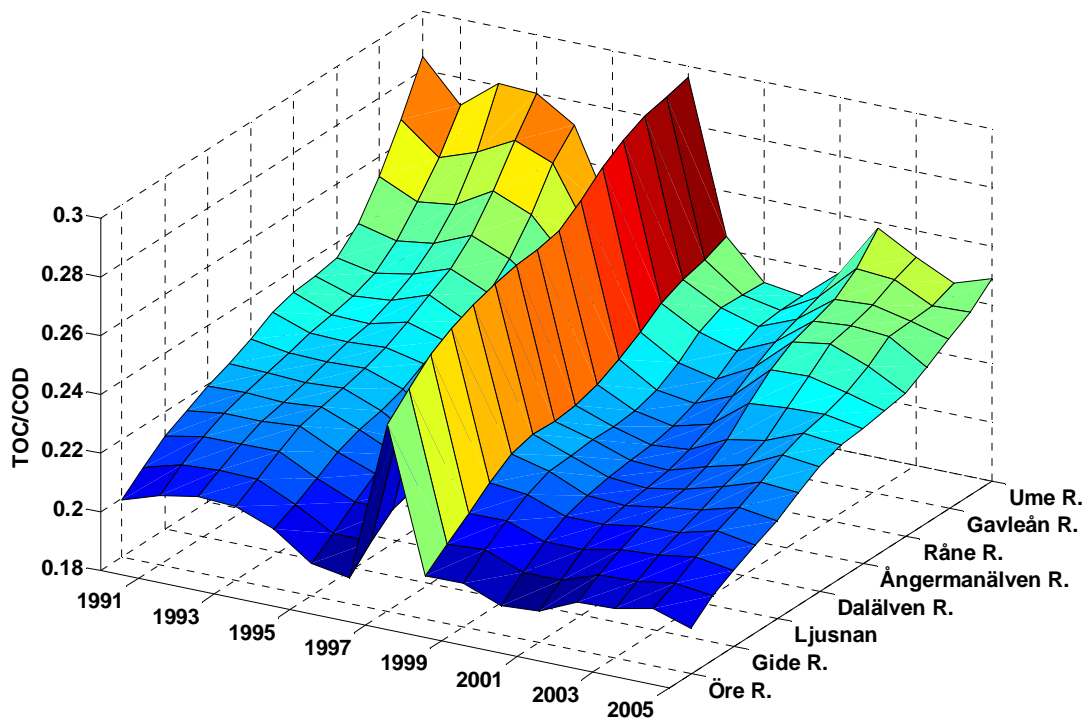
Since 1987, the amount of organic matter in Swedish surface waters has been measured as both TOC and COD (analysis of the latter using potassium permanganate as oxidant). Although there is no fixed relationship between the results obtained by the two methods, the data for each water body are normally strongly correlated, which makes it possible to identify time periods when the TOC or COD measurements have been biased. We chose to examine data from 1990 to 2005, because the first few years of TOC measurements were deemed to be less accurate.

Figure 9 illustrates the variation in TOC-to-COD ratios for fifteen major rivers in northern Sweden. This dataset was analysed using a model with two level shifts that were of the same size but had different signs. The timing of the level shifts was estimated from

the data, and Figure 10 shows the sum of the estimated smooth trend surface and level shifts. As expected, the algorithm identified 1997 as a period during which the data deviated strongly, and closer analysis showed a level shift of 0.062 for that year, with a standard error of 0.0038.

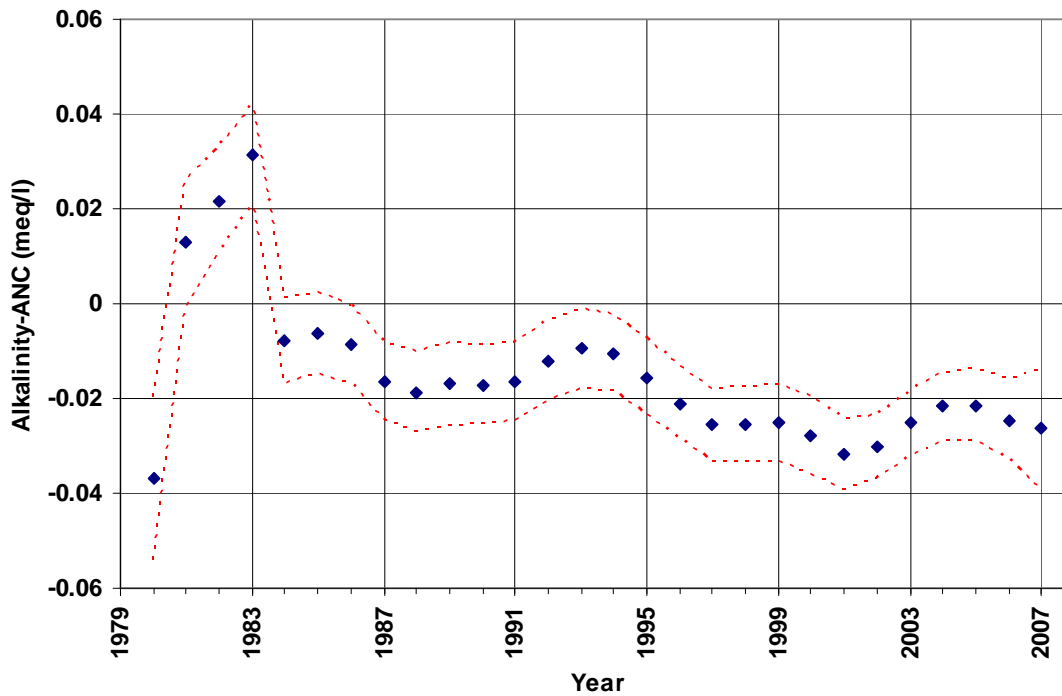


**Figure 9.** TOC-to-COD ratios recorded at the mouths of fifteen major rivers in northern Sweden. Monthly sampling was done in all the rivers throughout the investigated period.



**Figure 10.** Trend surface with discontinuities fitted to the data given in Figure 9. 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.

In our previously cited article on data quality (Wahlin and Grimvall 2008a), we claimed that alkalinity trends in Swedish groundwaters were contaminated by systematic measurement errors in the early 1980s. We reanalysed that dataset in the present work. 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). We found that our algorithm, which can accommodate observations that are unevenly distributed in time and space, confirmed our previous suspicion. Figure 11 illustrates how introduction of a new analytical procedure stabilized the annual mean of the estimated trend surface (including discontinuities) after 1984.



**Figure 11.** Annual means of trend levels (including discontinuities) fitted to differences between alkalinity and ANC in low-ANC samples from 77 Swedish groundwater sites.

## ***Discussion***

This article has demonstrated how smooth trends in a vector time series can be separated from abrupt level shifts that occur simultaneously in all coordinates. Such methods are obviously needed in environmental monitoring, but they can also be applied in almost any context in which several time series with similar trends are recorded. We developed our method primarily to facilitate unprejudiced searches for abrupt level shifts at unknown time points. However, our procedure is still applicable if we know when there has been some kind of major change, such as a switch in laboratory, personnel, analytical procedure, or sampling technique. More specifically, we can test the statistical significance of a level shift and examine whether the step size increases or decreases with the coordinate of the analysed vector time series.

The greatest strength of our method is its adaptive character. If the analysed time series have different trends, it is unlikely that the mean function can be made stepwise constant

by subtracting a suitable reference from each series. This implies that, in such cases, none of the existing methods mentioned in the introduction are applicable. It has been suggested that ordinary regression models in which the mean function is linear between the change points can serve as alternatives to models with stepwise constant means (Alexandersson and Moberg 1997; Easterling and Peterson 1995). However, that model class forces the user to choose between constant and discontinuous trend slopes. Our method is based on the more natural assumption that the trend slope (after removing the level shifts at the change points) varies smoothly over the entire study period, and the selection of smoothing factors by cross-validation automatically adapts the degree of smoothness to the analysed data.

The limitations of our method are also related to its adaptive character. In principle, our technique can be generalized to handle multiple change points that occur at different times in different coordinates of the studied vector time series. However, there are two major obstacles to such generalizations. First, it is difficult to distinguish between smooth changes in the trend surface and the combined effect of multiple discontinuities, which occur relatively close in time. In addition, there are computational obstacles to the handling of multiple change points. The model we propose is a three-step back-fitting algorithm in which the smooth trend surface, the regression coefficients of the covariates, and the discontinuities are estimated separately. In this type of algorithm, each step must be very fast, because it is repeated many times during the model fitting and an even larger number of times during the cross-validation and the analysis of resampled data. Consequently, it is not feasible to make unprejudiced searches for complex patterns of discontinuities in the presence of smooth trends that may vary from coordinate to coordinate.

Some comments should also be made about the resampling technique we used to assess the uncertainty of the detected level shifts. Our technique offers the important advantage of taking into account the correlation structure of the model residuals. Moreover, it is well coordinated with the smoothers used to extract the trend surface. However, like any other form of residual resampling, our method creates a new resampled dataset by adding

resampled residuals to fitted response values. Consequently, it is tacitly assumed that the errors in the fitted responses are considerably smaller than the individual error terms. This assumption is reasonable as long as the fitted responses are influenced by a large number of observations, but it is less appropriate if there are only a few influential data points. In practice, this implies that the uncertainty estimates are reliable for models with relatively strongly regularized trend surfaces (large  $\lambda$ -values).

### ***Acknowledgements***

The authors are grateful for financial support from the Swedish Environmental Protection Agency. The Geological Survey of Sweden funded the case studies of groundwater quality.

### ***References***

Aguilar E., Auer I., Brunet M., Peterson T.C. and Wieringa J. (2003). Guidelines on climate metadata and homogenization. WMO TD No. 1186, World Meteorological Organization.

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

Alexandersson H. and Moberg A. (1997). Homogenization of Swedish temperature data. Part I: homogeneity test for linear trends. *International Journal of Climatology*, 17, 25-34.

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

Easterling D.R. and Peterson T.C. (1995). A new method for detecting and adjusting for undocumented discontinuities in climatological time series. *International Journal of Climatology*, 15, 369-377.

Grath J., Ward R. and Quevauviller P. (eds) (2007). Common implementation strategy for the water framework directive. Guidance on groundwater monitoring. Office for Official Publications of the European Communities: Luxembourg.

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

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

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.

Jones P.D. (1995). The instrumental data record: its accuracy and use in attempts to identify the “CO<sub>2</sub>” signal. In von Storch H. and Navarra H. (eds), *Analysis of Climate Variability*. Springer Verlag: Berlin.

Klein Tank A.M.G., Wijngaard J.B., Können G.P., Böhm R., Demarée G., Gocheva A., Mileta M., Paschiardis S., Hejkrlik L., Kern-Hansen C., Heino R., Bessemoulin P., Müller-Westermeier G., Tzanakou M., Szalai S., Pálsdóttir T., Fitzgerald D., Rubin S., Capaldo M., Maugeri M., Leitass A., Bukantis A., Aberfeld R., van Engelen A.F.V., Førland E., Miletus M., Coelho F., Mares C., Razuvaev V., Nieplova E., Cegnar T., López A.J., Dahlström B., Moberg A., Kirchhofer W., Ceylan A., Pachaliuk O., Alexander L.V. and Petrovic P. (2002). Daily dataset of 20th-century surface air temperature and precipitation observations for European Climate Assessment. *International Journal of Climatology*, 22, 1441–1453.

Libiseller C., Grimvall A., Waldén J. and Saari H. (2005). Meteorological normalisation and non-parametric smoothing for quality assessment and trend analysis of tropospheric ozone data. *Environmental Monitoring and Assessment*, 100, 33-52.

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.

Picard F., Lebarbier E., Budinska E. and Robin S. (2007). Joint segmentation of multivariate Gaussian processes using mixed linear models. *Statistics for Systems Biology Group Research Report No. 5*, Jouy-en-Josas/Paris/Evry, INRA, France.

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

SLU (Swedish University of Agricultural Sciences) (2008). <http://www.ma.slu.se>. Accessed 2008-08-20.

Srivastava 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.

Szentimrey T. (1997). Statistical procedure for joint homogenization of climatic time series. *Proceedings of the 1st Seminar on Homogenization of Surface Climatological Data*, Budapest, Hungary, pp. 47-62.

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

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

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.