Semiparametric smoothers for trend assessment of multiple time series of environmental quality data

Grimvall, A., Hussian, M. and Libiseller, C.
Department of Mathematics, Linköping University, SE-58183 Linköping
e-mail:angri@mai.liu.se

Abstract

Multiple time series of environmental quality data with similar, but not necessarily identical, trends call for multivariate methods for trend detection and adjustment for covariates. Here, we show how an additive model in which the multivariate trend function is specified in a nonparametric fashion (and the adjustment for covariates is based on a parametric expression) can be used to estimate how the human impact on an ecosystem varies with time and across components of the observed vector time series. More specifically, we demonstrate how a roughness penalty approach can be utilized to impose different types of smoothness on the function surface that describes trends in environmental quality as a function of time and vector component. Compared to other tools used for this purpose, such as Gaussian smoothers and thin plate splines, an advantage of our approach is that the smoothing pattern can easily be tailored to different types of relationships between the vector components. We give explicit roughness penalty expressions for data collected over several seasons or representing several classes on a linear or circular scale. In addition, we define a general separable smoothing technique.

1 Introduction

Deterioration of an ecosystem is usually a slow process, and mathematical functions that describe the impact of humans on the environment will presumably vary smoothly over time. Furthermore, in many cases, a substantial fraction of the temporal variability in the measured state of the environment can be attributed to random fluctuations in weather conditions or other types of natural changeability. Consequently, there is an obvious need for statistical methods that enable simultaneous extraction of smooth trends and adjustment for covariates.

The approaches most often used to detect trends in environmental quality in the presence of covariates have been reviewed by Thompson and coworkers (2001). Regression methods predominate and several investigators have employed nonparametric or semiparametric techniques because they enable unprejudiced inference about the shape of the trend curves (Shively & Sager, 1999; Gardner & Dorling, 2000 a,b; Stålnacke & Grimvall, 2001; Giannitrapani *et al.*, 2004, 2005). In particular, it has been emphasized that generalized additive

models (GAMs) provide a suitable framework for such inference (Giannitrapani *et al*., 2004, 2005).

Our research group has focused on applications in which the collected data represent several classes of observations, and the models are estimated using a roughness penalty approach that allows the smoothness conditions to be selected just as carefully as other model features. Two papers have addressed time series of riverine load data collected over several seasons (Stålnacke & Grimvall, 2001; Hussian *et al*., 2004), and Libiseller and Grimvall (2005) have recently presented a method for trend analysis and normalization of atmospheric deposition data representing several wind sectors. Both the riverine loads by season and the deposition by wind sector can be regarded as vector time series of annual data for which a joint analysis of temporal trends would be desirable. Additional examples of such multivariate data are observations from several sampling sites along a gradient or several congeners of an organic pollutant.

Here, we show how an additive model in which the multivariate trend function is specified in a nonparametric fashion (and the adjustment for covariates is based on a parametric expression) can be used to estimate how the human impact on the ecosystem of interest varies with time and across components of the observed vector time series. More specifically, we demonstrate how a roughness penalty approach can be utilized to impose different types of smoothness on the function surface describing the trends in environmental quality as a function of time and vector component.

2 Basic semiparametric model

Let

$$\boldsymbol{y}_t = (y_t^{(1)}, ..., y_t^{(m)})', \quad t = 1, ..., n$$

denote an *m*-dimensional vector time series representing the observed state of the environment at *n* equidistant time points, and let

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

be a matrix that includes contemporaneous values of p explanatory vectors representing natural fluctuations in $\boldsymbol{y}_t$. Further, assume that

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

where the sequence of vectors $\boldsymbol{\alpha}_t = (\alpha_t^{(1)}, ..., \alpha_t^{(m)})', t = 1, ..., n,$ 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 independent of each other and of the explanatory variables.

In principle, there is no relationship between the m components of the vector time series $\boldsymbol{y}_t$ in formula 1, because all error terms are assumed to be statistically independent and each series has its own set of intercept and slope parameters. However, estimation of this semiparametric model requires that smoothness conditions be introduced to make the degrees of freedom (effective dimension) of the model less than the number of observations, which imposes constraints on the variation across components. In addition, we note that smoothness conditions for the intercept parameters $\alpha_t^{(j)}$ are introduced in a more natural manner, if we reformulate the model so that the intercepts represent the expected response values when the covariates are equal to their expectations, that is

$$y_t^{(j)} = \alpha_t^{(j)} + \sum_{i=1}^{p} \beta_i^{(j)} (x_{it}^{(j)} - E(x_{it}^{(j)})) + \varepsilon_t^{(j)}, \quad j = 1, ..., m, \quad t = 1, ..., n \tag{2}$$

The following discussion considers how the roughness penalty expressions can be tailored to achieve different smoothing patterns. The smoothing over time (years) is identical for all variants, whereas the smoothing across vector components differs. For example, when data represent different wind sectors, it is natural to introduce constraints that force the trend levels (intercepts) of adjacent sectors to be similar. Likewise, when data represent several seasons, it is natural to force the trend levels for the last season in one year and the first season in the following year to

be close to each other. We use the terms circular and sequential smoothing for these types of constraints on the intercept parameters (Figures 2.3a, b). Data collected along a gradient may also require tailored smoothing in two directions: over time and along the sampled gradient (Figure 2.3c).

a)

b)

c)

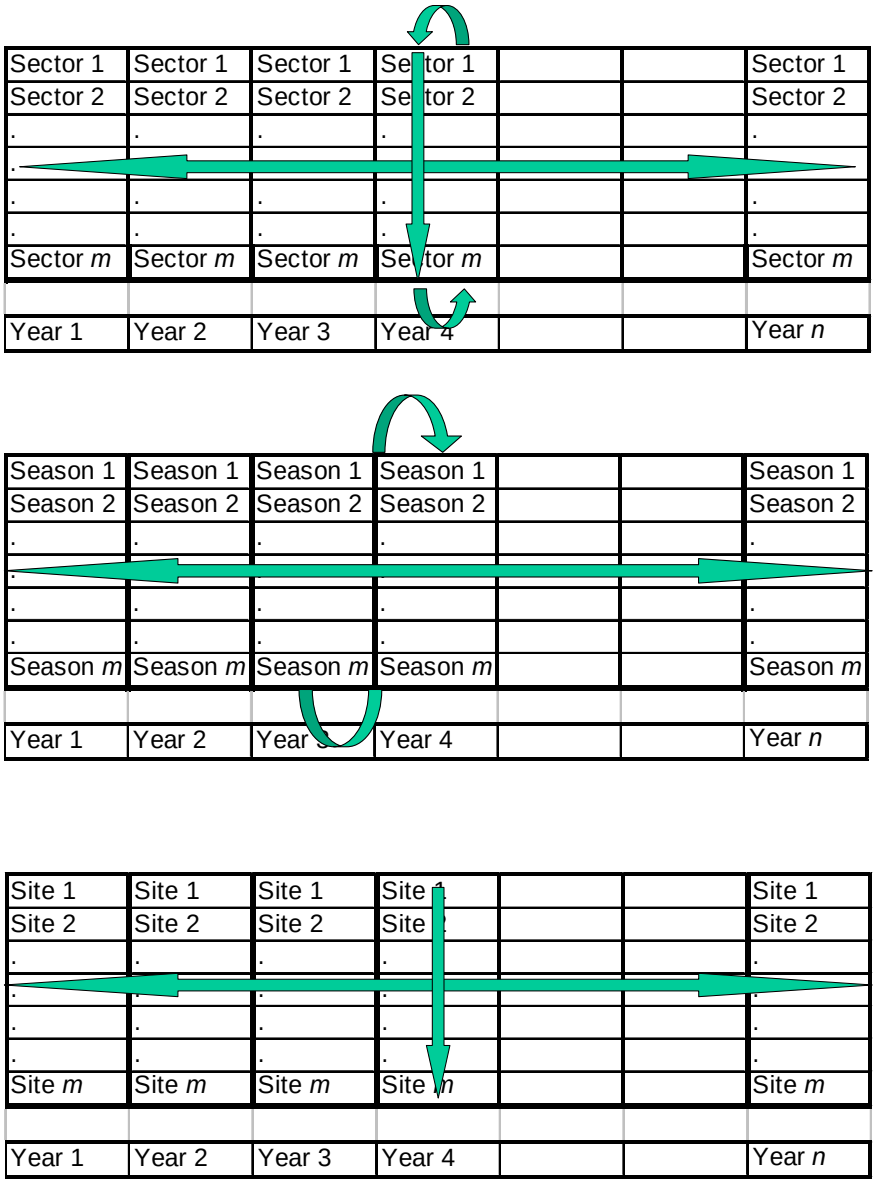

Figure 1. *Different types of smoothing patterns for the intercept of the basic semiparametric model. The three graphs show circular smoothing for data representing several sectors (a), sequential smoothing for data collected over several season (b), and gradient smoothing for data collected at different sites along a gradient (c).*

3 Circular smoothing

Circular smoothing is desirable when the collected data represent different classes on a circular scale. The previously mentioned example of deposition data for different wind sectors can be used to illustrate this situation. The expected responses are similar for adjacent classes, but the first and last classes are also closely interrelated.

To introduce circular smoothing in our semiparametric model, we estimate the parameters by minimizing the sum

$$S_0(\boldsymbol{\alpha},\boldsymbol{\beta}) + \lambda_1 S_1(\boldsymbol{\alpha}) + \lambda_2 S_2(\boldsymbol{\alpha}) \tag{3}$$

where λ_1 and λ_1 are roughness penalty factors,

$$S_0(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{t=1}^{n}\sum_{j=1}^{m}(y_t^{(j)} - \alpha_t^{(j)} - \sum_{i=1}^{p}\beta_i^{(j)}(x_{it}^{(j)} - \bar{x}_{i\cdot}^{(j)}))^2 \tag{4}$$

is the residual sum of squares,

$$S_1(\boldsymbol{\alpha}) = \sum_{t=2}^{n-1}\sum_{j=1}^{m}(\alpha_t^{(j)} - \frac{\alpha_{t-1}^{(j)} + \alpha_{t+1}^{(j)}}{2})^2 \tag{5}$$

represents smoothing of the multivariate temporal trend over time, and

$$S_2(\boldsymbol{\alpha}) = \sum_{t=1}^{n}\sum_{j=1}^{m}(\alpha_t^{(j)} - \frac{\alpha_t^{(j-1)} + \alpha_t^{(j+1)}}{2})^2 \tag{6}$$

stands for smoothing across components of this trend. As usual, $\bar{x}_{i\cdot}$ depicts a mean value and, to simplify the notation for the circular smoothing, we introduce the symbols

$$\alpha_t^{(m+1)} = \alpha_t^{(1)} \text{ and } \alpha_t^{(0)} = \alpha_t^{(m)}$$

4 Sequential smoothing

When data are collected over several seasons, there is an obvious relationship between the observations for adjacent seasons. This feature can be incorporated into the smoothing pattern by letting $s = t(m-1) + j$ represent the sequential order of the observations and setting

$$S_2(\boldsymbol{\alpha}) = \sum_{s=2}^{mn-1}(\alpha_s - \frac{\alpha_{s-1} + \alpha_{s+1}}{2})^2 \tag{7}$$

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

5 Gradient smoothing

Gradient smoothing can be suitable if the collected data come from sampling sites located along a transect or along a gradient of elevation, temperature or precipitation. Such smoothing can also be appropriate for data representing measured concentrations of chemical compounds, which can be ordered linearly with respect to, for instance, volatility, polarity, or lipophilicity. Regardless of how the linear ordering is defined, smoothing across coordinates can be accomplished by setting

$$S_2(\boldsymbol{\alpha}) = \sum_{t=1}^{n} \sum_{j=2}^{m-1} (\alpha_t^{(j)} - \frac{\alpha_t^{(j-1)} + \alpha_t^{(j+1)}}{2})^2 \qquad (8)$$

6 Separable smoothing

The circular and gradient smoothing are special cases of more general smoothing patterns that can be defined by setting

$$S_2(\boldsymbol{\alpha}) = \sum_{t=1}^{n} \sum_{(j_1, j_2, j_3) \in W} w(j_1, j_2 . j_3)(\alpha_t^{(j_1)} - \frac{\alpha_t^{(j_2)} + \alpha_t^{(j_3)}}{2})^2 \qquad (9)$$

where W comprises all triplets in which the coordinates are distinct integers between 1 and m, and the (nonnegative) weight $w(j_1, j_2, j_3)$ represents a measure of the proximity of the coordinate j_1 to the coordinates j_2 and j_3. We refer to this as separable smoothing, because the expression that is minimized during the parameter estimation contains a weighted sum of two roughness penalty terms, S_1 and S_2, which represent smoothing in two different directions (over time and across coordinates, respectively). The inverse Euclidean distance may be a suitable measure of proximity when the analyzed vector time series of data come from measurements made at permanent plots in a sampling area or at different stations in a monitoring program. However, any proximity measure in any space can be used to define the weights w in formula 9.

7 Smoothing and normalization

The models discussed in this article can be used to produce two different types of outputs. First, we can estimate the temporal trend $\alpha_t^{(j)},\ j = 1, ..., m, t = 1, ..., n,$ by suppressing all types of random variation in the collected data. Secondly, we can normalize the observed data by removing the variation that can be attributed to the covariates, that is, by forming

$$\tilde{y}_t^{(j)} = y_t^{(j)} - \sum_{i=1}^{p} \hat{\beta}_i^{(j)} (x_{it}^{(j)} - \bar{x}_i^{(j)}), \quad j = 1, ..., m, \quad t = 1, ..., n \tag{10}$$

where $\hat{\beta}_i^{(j)}$ is the estimated regression coefficient for the *j*th component of the *i*th covariate. Figure 2 illustrates the results obtained when gradient smoothing was used to summarize the trends in mean annual concentrations of mercury in muscle tissue from flounder (*Platichthys flesus*) caught in the German Bight at five stations located at varying distances from the mouth of the Elbe River, which was heavily polluted until the beginning of the 1990s. Figure 3 illustrates the difference between the smooth trend surface and the considerably rougher surface of normalized values for monthly loads of nitrogen carried by the Rhine River.

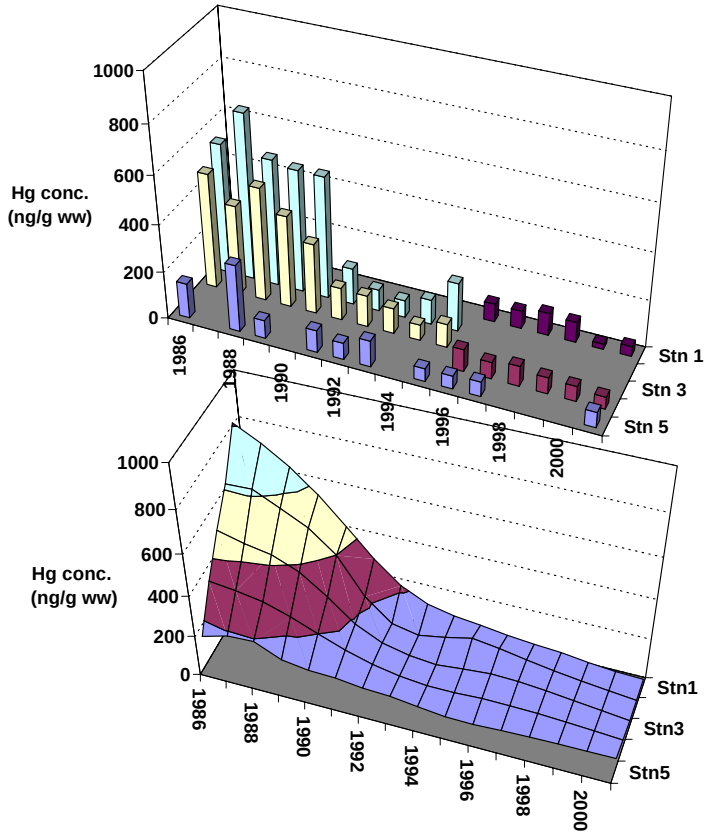

Figure 2. *Trend assessment of the concentration of mercury in muscle tissue from flounder (Platichthys flesus) caught in the German Bight at five stations (53° 53' N, 9° 11' E; 53° 52' N, 8° 52' E; 53° 56' N, 8° 38' E; 53° 57' N, 8° 30' E; 53° 45' N, 8° 2' E) in or outside the mouth of the Elbe River. The two graphs show observed annual means and a trend surface obtained by gradient smoothing ($\lambda_1 = 2$, $\lambda_2 = 0.5$).*

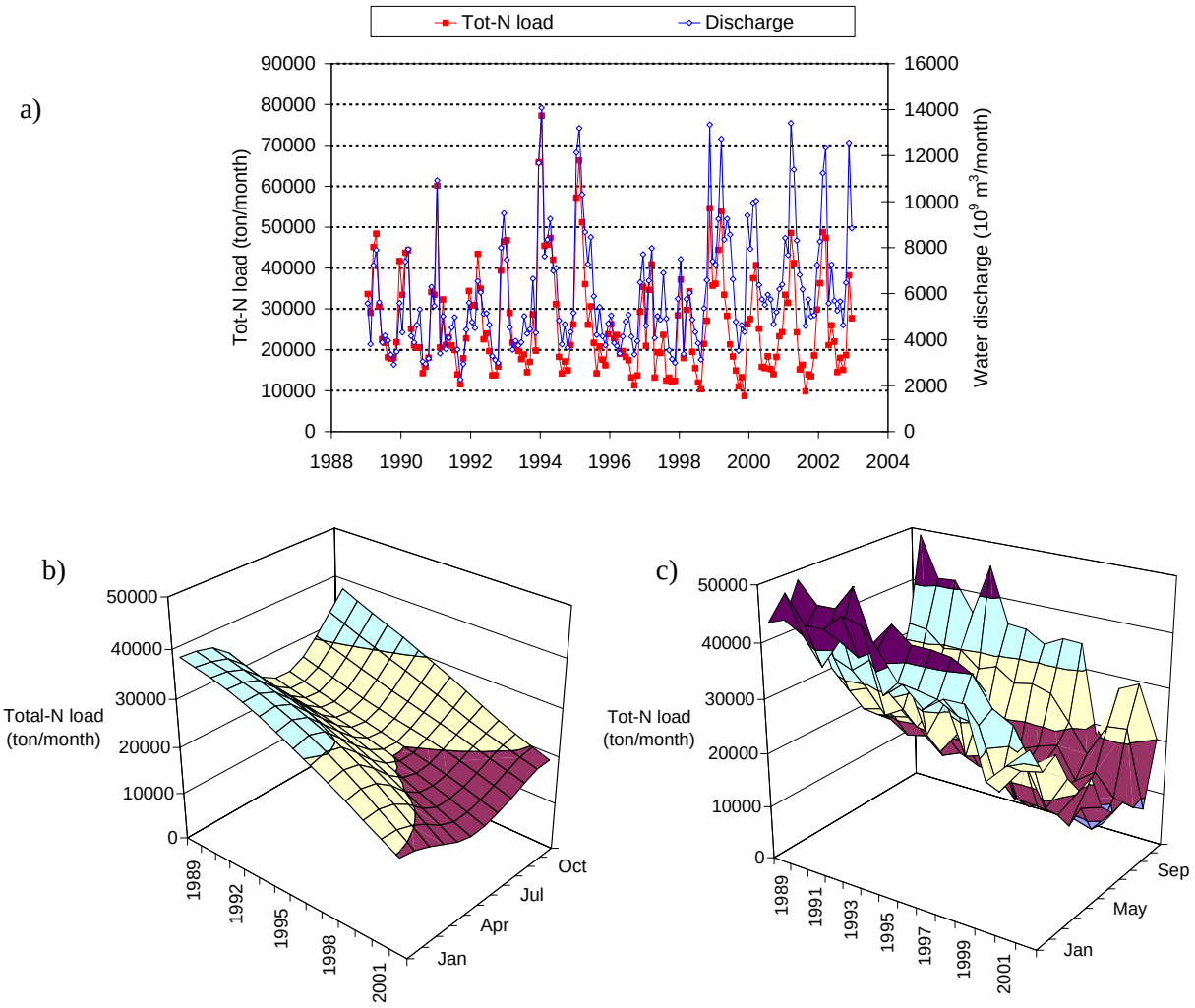

Figure 3. *Trend assessment of the total nitrogen load in the Rhine River at Lobith on the border between Germany and The Netherlands. The three graphs show observed monthly nitrogen loads and water discharge values (a), the estimated trend function (b), and flow-normalized monthly loads (c).*

8 Computational aspects

For fixed values of the smoothing factors λ_1 and λ_2, the semiparametric models described in this article can be estimated by using back-fitting algorithms in which estimation of the slope parameters for fixed intercepts is alternated with estimation of the intercepts for fixed slopes. More specifically, ordinary regression algorithms can be used to estimate the slope parameters, whereas a system of linear equations with *nm* unknowns must be solved to estimate the intercepts. The latter can be achieved by Cholesky factorization of the coefficient matrix and sequential determination of the unknowns (Silverman & Green, 1994). In the case of sequential smoothing,

Stålnacke and Grimvall (2001) demonstrated that the coefficient matrix is a positive definite symmetric band matrix with lower and upper bandwidth *2m*. The other smoothing techniques described here give rise to symmetric band matrices that have the same bandwidth. Also, it can be noted that, for a given bandwidth, the computational burden is proportional to the number of time points *n*.

The smoothing factors λ_1 and λ_2 that control the effective dimension of the estimated models can be determined by cross-validation. However, it should be noted that conventional leave-one-out cross-validation may lead to over-fitting if the error terms are correlated (Shao, 1993; Libiseller & Grimvall, 2003). Hence, we propose block cross-validation in which the observed response values are divided into *n* blocks, each composed of *p* contemporaneous components of $\boldsymbol{y}_t$.

9 Further generalizations

9.1 *Time-varying slope parameters*

The model introduced in formula 2 has time-dependent intercepts, whereas the slope parameters are not dependent on time. Models in which the intercepts are time independent and the time-dependence of the slope parameters is controlled by roughness penalty expressions can be specified and estimated in a similar manner (Stålnacke & Grimvall, 2001). However, for computational reasons, it is not feasible to handle more than one explanatory variable in such semiparametric models. Furthermore, it is not practicable to let both the intercept and the slope parameters vary with time due to following: back-fitting algorithms for such models may have convergence problems, and the computational burden increases dramatically with the number of roughness penalty factors determined by cross-validation.

9.2 *Missing and multiple values*

To keep the notation as simple as possible, thus far we have assumed that we have exactly one observation of the response variable and the *p* explanatory variables for each combination of time point (year) and vector component. As pointed out by Hussian *et al*. (2004), the algorithms used to estimate the semiparametric models can easily be adapted to accommodate missing values by changing the diagonal elements of the above-mentioned band matrix. More generally, our approach can handle data sets of the type

$$(y_{t(k)}^{j(k)}, x_{1,t(k)}^{j(k)}, ..., x_{p,t(k)}^{j(k)}), \quad k = 1, ..., N$$

where $t(k)$ and $j(k)$ respectively denote the time point and the vector component of the kth observation.

9.3 *Nonparametric analogues*

The nonparametric specification of the intercepts can be justified by a desire to make unprejudiced inference about the shape of the trend surface. Similar arguments might rationalize the use of models in which both the trend and the influence of covariates are specified nonparametrically, for example, those of the type

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

where $h_i^{(j)}, i = 1, ..., p, j = 1, ..., m,$ are smooth functions. Furthermore, this generalization is easy to program when a back-fitting algorithm is used to estimate the model parameters, because, with such an approach, the estimation of $h_i^{(j)}, i = 1, ..., p, j = 1, ..., m,$ for given intercepts can be reduced to mp nonparametric univariate regressions. However, it should be noted that models with very few structural constraints may lead to problems with over-fitting and slow convergence of the back-fitting algorithm.

10 Discussion

The current techniques for detecting trends in time series of environmental quality data are dominated by approaches in which the collected data are analyzed separately for each site and response variable. The major exceptions to this rule are the widespread use of multivariate linear models and the growing interest in spatio-temporal geostatistical models. In this article, we have demonstrated that roughness penalty approaches to the estimation of semiparametric regression models constitute a very flexible group of techniques for simultaneous detection of temporal trends and adjustment for covariates in vector time series of environmental quality data.

In contrast to ARIMAX models and other classical multivariate time series models, which are based on stationarity or very simple forms of nonstationarity, our method enables careful modeling of the nonstationary features of the collected data. This property, it shares with a large class of methods that can be referred to as smoothing techniques for response surfaces. Thin plate splines, vector generalized additive models (GAMs), and kernel smoothers (Wahba, 1990; Yee & Wild, 1996; Härdle, 1997; Hastie *et al.*, 2001) can all be appropriate for estimating nonlinear trends. However, this article shows that our roughness penalty approach has the advantage that

the smoothing pattern can be tailored to take into account almost any relationship between the different components of the observed random vectors. Furthermore, the degree of smoothing can be fine-tuned in two directions without making the computational burden insurmountable.

Dynamic factor analysis (DFA) represents another promising approach to assessment of trends in vector time series data (Zuur *et al*., 2003). Inasmuch as DFA is a latent variable technique, it is particularly useful when the dimension of the analyzed vector time series is high. Nonetheless, it might be worth considering our method for such data. This is particularly true when then there is considerable prior knowledge about the interrelationship between the vector components or the proximity of the vector components can be modeled.

Finally, it is worth noting that, although some of the computational aspects of parameter estimation in semiparametric models may require special attention (Schimek, 2001), the basic principles of such models are easy to communicate to a wide audience.

11 Acknowledgements

The authors are grateful for financial support from the Swedish Environmental Protection Agency.

12 References

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

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

Giannitrapani M., Scott M., Bowman A., and Smith R. (2004). *A statistical method to analyze trends in EMEP data*. EMEP Assessment Report (www.emep.int/assessment/appendix1.pdf).

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

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

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

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

Libiseller C. and Grimvall A. (2005). *Trend analysis of CAMP data regarding wet and dry nitrogen deposition*. Report to the 2005 meeting of the OSPAR working group on inputs to the marine environment (INPUT). London, OSPAR.

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

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

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

Silverman B.W. and Green P.J. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach.* London, Chapman & Hall.

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

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

Wahba G. (1990). *Spline models for observational data*. Philadelphia, SIAM.

Yee T.W. and Wild C.J. (1996) Additive extensions to generalized estimation equation methods. *Journal of the Royal Statistical Society B*, **58**, 711-725.

Zuur A.F., Fryer R.J., Jolliffe I.T., Dekker R., and Beukema J.J. (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, *7*, 665-685