European Congress on Computational Methods in Applied Sciences and Engineering
ECCOMAS 2004
P. Neittaanmäki, T. Rossi, K. Majava, and O. Pironneau (eds.)
V. Capasso and W. Jäger (assoc. eds.)
Jyväskylä, 24-28 July 2004

MONOTONIC REGRESSION FOR ASSESSMENT OF TRENDS IN ENVIRONMENTAL QUALITY DATA

Mohamed Hussian[1], Anders Grimvall[1], Oleg Burdakov[1], and Oleg Sysoev[2]

[1]Department of Mathematics, Linköping University, SE-58183 Linköping, Sweden
E-mail: mohus@mai.liu.se, angri@mai.liu.se, olbur@mai.liu.se
Web site: www.mai.liu.se

[2]Faculty of Control and Applied Mathematics, Moscow Institute of Physics and Technology,
Institutskij per. 9, Dolgoprudnyj, Moscow region 141700, Russia
E-mail: osysoev@mail.ru

Keywords: Monotonic regression, Response surface, Time series decomposition, Normalisation

Abstract. *Monotonic regression is a non-parametric method that is designed especially for applications in which the expected value of a response variable increases or decreases in one or more explanatory variables. Here, we show how the recently developed generalised pool-adjacent-violators (GPAV) algorithm can greatly facilitate the assessment of trends in time series of environmental quality data. In particular, we present new methods for simultaneous extraction of a monotonic trend and seasonal components and for normalisation of environmental quality data that are influenced by random variation in weather conditions or other forms of natural variability. The general aim of normalisation is to clarify the human impact on the environment by suppressing irrelevant variation in the collected data. Our method is designed for applications that satisfy the following conditions: (i) the response variable under consideration is a monotonic function of one or more covariates; (ii) the anthropogenic temporal trend is either increasing or decreasing; (iii) the seasonal variation over a year can be defined by one increasing and one decreasing function. Theoretical descriptions of our methodology are accompanied by examples of trend assessments of water quality data and normalisation of the mercury concentration in cod muscle in relation to the length of the analysed fish.*

1. INTRODUCTION

Monotonic responses and relationships are widespread in all types of environmental systems. For example, it is common that the rates of chemical and microbial processes increase with temperature. Also, the concentrations of many contaminants in living organisms increase with the age or size of the analysed individual, and fluxes of substances through terrestrial and aquatic systems can increase with the amount and intensity of precipitation. The simplest forms of monotonic relationships can easily be described by using an appropriate parametric model, and numerous algorithms have been developed to fit such models to observed data. However, more complex relationships involving two or more explanatory variables can require non-parametric modelling. This is especially true if the response includes a threshold effect or is strongly non-linear in some other respect.

Monotonic regression is a non-parametric method that is designed for applications where the expected value of a response variable (y) increases or decreases in one or more explanatory variables ($x_1, \ldots, x_p$). The most commonly used computational method for this type of regression is the so-called pool-adjacent-violators (PAV) algorithm [1, 2, 3]. When $p = 1$, this algorithm is computationally efficient, and it provides solutions that are optimal in the sense that the mean square error is minimised. When $p > 1$, the PAV algorithm has proven useful for estimating monotonic responses to explanatory variables that are varied at only a few levels [4, 5, 6, 7]. However, it was not until Burdakov and colleagues [8, 9] recently generalised the PAV algorithm from fully to partially ordered data that it became feasible to handle typical regression data that include one or more continuous variables. The cited reports also explain the ways in which the generalised pool-adjacent-violators (GPAV) algorithm is superior to currently used algorithms that are based on simple averaging techniques [10, 11, 12] or quadratic programming [13, 14].

Here, we show that the GPAV algorithm has important applications in several areas of environmental science and management. In particular, we illustrate how this algorithm can be used in the following contexts:

(i) estimation of response surfaces that are known to be monotonic in two or more variables;

(ii) simultaneous extraction of seasonal components and a monotonic trend from a univariate time series;

(iii) normalisation of time series of environmental quality data.

The first of these tasks is also highly relevant in many areas other than environmental science; for example, monotonic regression is often appropriate for estimating dose-response curves in experimental studies [6, 7]. The second task entails time series decomposition, which is a classical undertaking in official statistics (e.g., [15]). The method we present takes into account that many seasonal patterns in the environment can be decomposed into one increasing and one decreasing phase. The third task, normalisation or adjustment, aims to

clarify the human impact on the environment by removing weather-dependent fluctuations or other natural variability in the collected data [16].

2. ESTIMATION OF A MONOTONIC RESPONSE IN TWO OR MORE EXPLANATORY VARIABLES

As we already pointed out, the GPAV algorithm is particularly useful when the expected response is monotonic in two or more explanatory variables and at least one of these variables is continuous. Such situations arise naturally when evaluating time series of environmental quality data for temporal trends. First, interest is often focused on monotonic trends. Second, almost all measurements of the state of the environment are influenced by weather conditions or other covariates, and it is more the rule than the exception that the relationships between the response variable and the covariates under consideration are monotonic.

Figure 1 (a and b) illustrates how monotonic regression can be used to describe data on the concentration of mercury in Atlantic cod (*Gadu morhua*) in relation to sampling year and body length. The increase in mercury with increasing size of the fish is obvious in the two diagrams. In addition, the response surface in Figure 1b indicates a downward temporal trend.

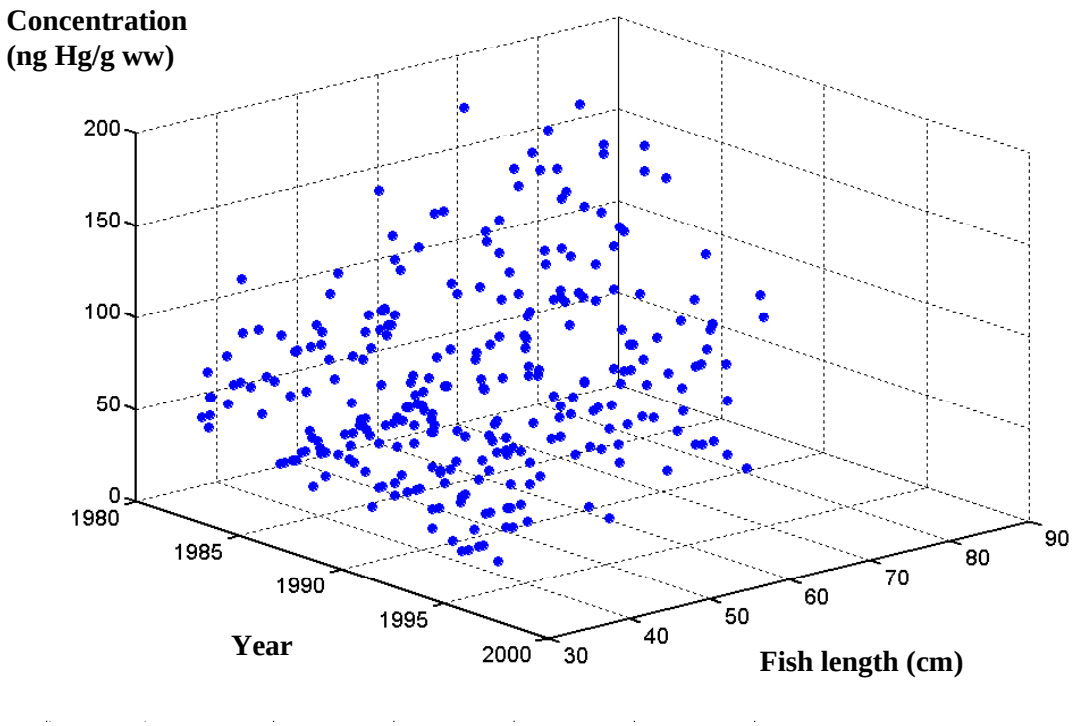


Figure 1a. Concentration of mercury in muscle tissue from Atlantic cod (*Gadu morhua*) caught in the North Sea (53° 10' N, 2° 5' E). The data represent observed concentrations (ng Hg/g ww) in relation to sampling year and body length length of the analysed fish.

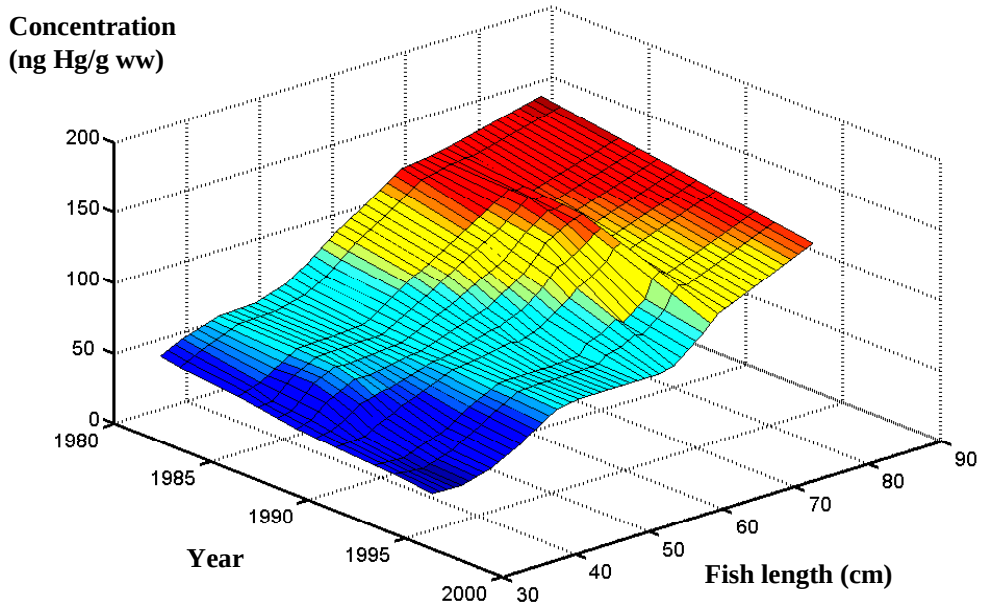


Figure 1b. Concentration of mercury in muscle tissue from Atlantic cod (*Gadu morhua*) caught in the North Sea (53° 10' N, 2° 5' E). The response surface was obtained by first using the GPAV algorithm for monotonic regression and then employing locally weighted scatter-plot smoothing to extrapolate the fitted regression values to a fine grid (see also section 4.3).

3. SIMULTANEOUS ESTIMATION OF A MONOTONIC TREND AND SEASONAL EFFECTS

When data are collected over several seasons, monotonic regression models may appear to be inadequate. However, many seasonal patterns can be decomposed into increasing and decreasing phases, and this enables the use of various approaches based on monotonic regression. If we let $y_1, y_2, \ldots, y_n$ denote a time series of data collected over m seasons, and let $\hat{y}_i$ denote the sum of the trend and seasonal components at time i, it is possible to determine $\hat{y}_i$ by minimising

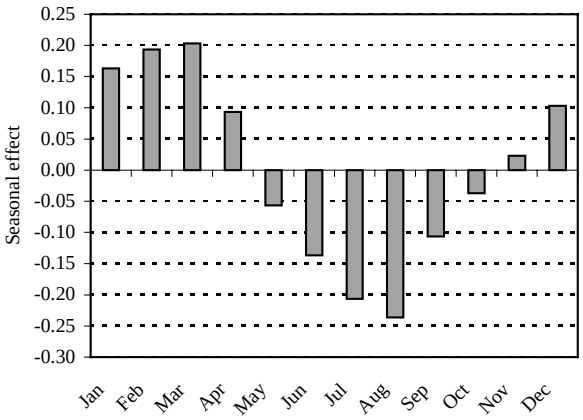
$$S = \sum_i (y_i - \hat{y}_i)^2$$

under a set of simple constraints, and we can also introduce these constraints by employing a monotonic regression model.

Let us, for the sake of clarity, consider a seasonal pattern of the type illustrated in Figure 2. Let us also assume that we would like to extract a non-increasing trend function from the collected data. If we then perform a monotonic regression using sampling year and the variables x_1 and x_2 as explanatory variables, the fitted values $\hat{y}_i$ must be non-increasing for each season, i.e.,

$$\hat{y}_i \geq \hat{y}_{i+m}, \quad i = 1,\ldots,n-m.$$

In addition, the fitted values representing different seasons in the same year must have non-increasing and non-decreasing phases with the same duration as in Figure 2.



Month	x_1	x_2
Jan	0	0
Feb	1	0
Mar	2	0
Apr	2	-1
May	2	-2
Jun	2	-3
Jul	2	-4
Aug	2	-5
Sep	3	-5
Oct	4	-5
Nov	5	-5
Dec	6	-5

Figure 2. Seasonal pattern comprising increasing and decreasing phases, and a possible coding of these phases.

Figure 3 (a and b) illustrates a set of monthly flow-weighted concentrations of total nitrogen in the Elbe River and the monotonic trend and seasonal components that could be extracted from these data. The goodness-of-fit to observed data reached a maximum when we let the seasonal effects have a maximum in March and a minimum in August. Furthermore, it can be noted that the downward trend was particularly strong after the reunification of Germany in 1990.

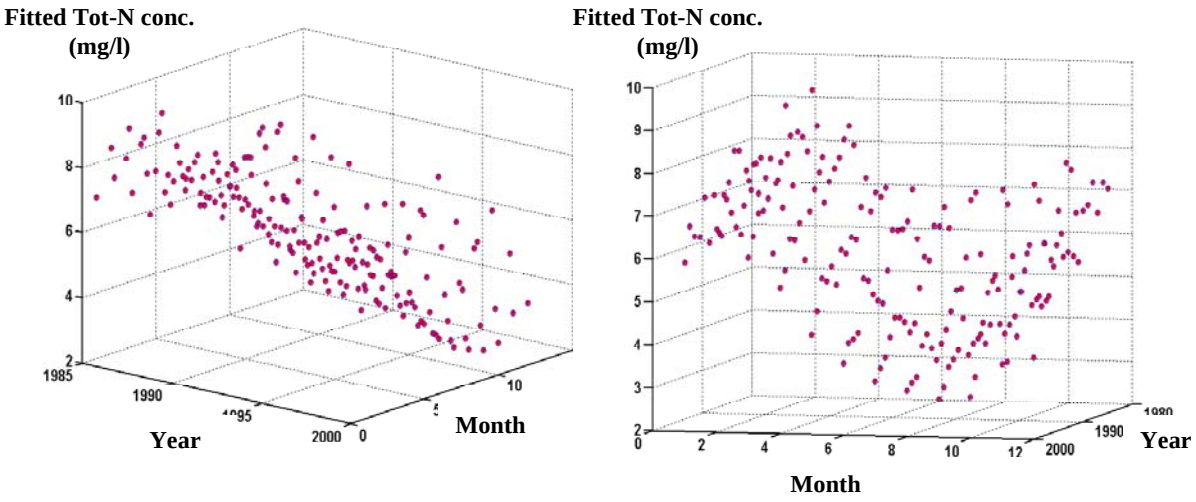


Figure 3a. Monthly mean concentrations of total nitrogen (Tot-N) measured in the Elbe River at Brunsbüttel downstream of Hamburg.

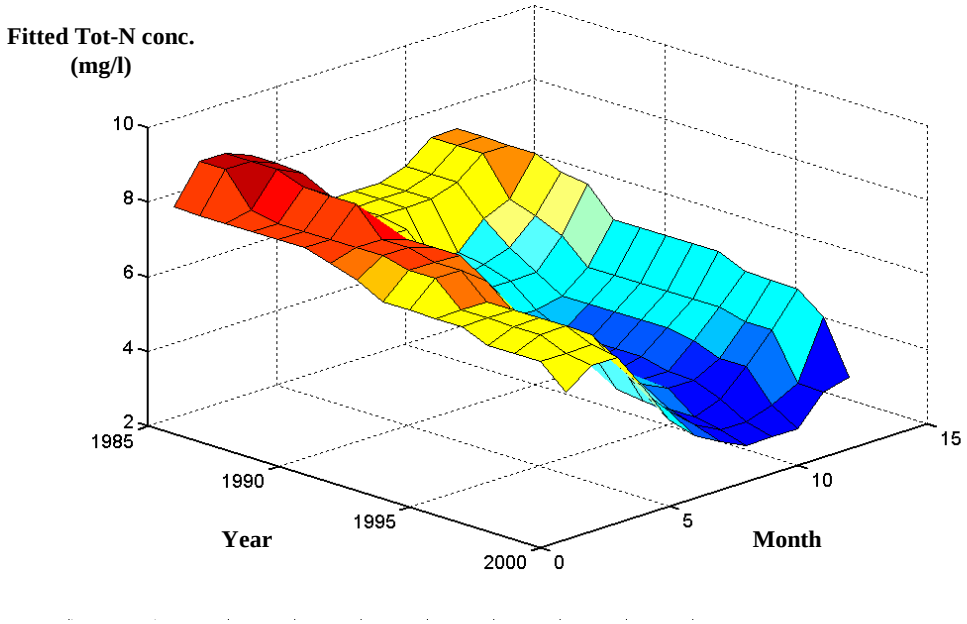


Figure 3b. Response surface obtained by applying monotonic regression to monthly mean concentrations of total nitrogen (Tot-N) measured in the Elbe River at Brunsbüttel downstream of Hamburg.

4. NORMALISATION OF TIME SERIES OF ENVIRONMENTAL QUALITY DATA

4.1 Normalisation formulae

The general aim of normalisation is to remove irrelevant variation in the collected data. The basic idea is simple. If observations of meteorological or other naturally fluctuating variables makes us believe that the studied response variable takes a value that is c units higher than the average response, then normalisation implies that we subtract this expected increase c from the observed response. A general probabilistic framework for normalisation was recently presented by Grimvall and co-workers [16]. Here, we discuss normalisation based on monotonic regression models.

Let us assume that the observed values of the response variable y have the general form

$$y_i = f(x_{1i}, \ldots, x_{pi}) + \varepsilon_i, \; i = 1, \ldots, n$$

where f is a deterministic function of p explanatory variables $x_1, \ldots, x_p$, and ε_i, $i = 1, \ldots, n$ depicts a sequence of independent, identically distributed random errors with mean zero. We can then normalise the observed responses with respect to $x_1, \ldots, x_q$ by forming

$$\tilde{y}_i = y_i - \left(\hat{f}(x_{1i}, \ldots, x_{qi}, x_{q+1,i}, \ldots, x_p) - \hat{f}(\bar{x}_{1i}, \ldots, \bar{x}_{qi}, x_{q+1,i}, \ldots, x_p) \right)$$

where $\hat{f}$ denotes an estimate of f and $(\bar{x}_{1i}, ..., \bar{x}_{qi}), i = 1,..., n$ is a sequence of given values of $x_1, \ldots, x_q$. Typically, these given values represent averages taken over the entire data set or subsets thereof. For example, if data are collected over several seasons, we can let $(\bar{x}_{1i}, ..., \bar{x}_{qi}), i = 1, ..., n$ represent seasonal means of $x_1, \ldots, x_q$.

Regardless of how the normalisation is carried out, we must be able to estimate the values of f at two sets of points: an estimation set

$$A = \left\{(x_{1i}, ..., x_{pi}), i = 1, ..., n\right\}$$

for which we have observed response values $\{y_i, i = 1, \ldots, n\}$, and an evaluation set

$$B = \left\{(\bar{x}_{1i}, ..., \bar{x}_{qi}, x_{q+1,i}, ..., x_{pi}), i = n+1, ..., n+m\right\}$$

for which no observations exist. The GPAV algorithm provides estimates of f for all points in the estimation set. It remains to extrapolate $\hat{f}$ to the evaluation set under the constraint that $\hat{f}$ is monotonic in each of the coordinates.

4.2 Extrapolation of monotonic responses to new points

Let $\chi = (\chi_1, \ldots, \chi_p)$ be a point to which $\hat{f}$ shall be extrapolated from a given estimation set A. We can then define two subsets of A. The first subset

$$L_\chi = \left\{(x_{1i}, ..., x_{pi}) \in A\,;\, x_{ki} \le \chi_k, k = 1, ..., p\right\}$$

contains all points in A that are dominated by χ. The second subset

$$U_\chi = \left\{(x_{1i}, ..., x_{pi}) \in A\,;\, x_{ki} \ge \chi_k, k = 1, ..., p\right\}$$

comprises all points in A that dominate χ.

Let us also for the moment assume that both L_χ and U_χ are nonempty. Then the expression

$$y_L = \max\left\{\hat{f}(x_{1i}, ..., x_{pi})\,;\, (x_{1i}, ..., x_{pi}) \in L_\chi\right\}$$

provides a lower limit for the values of $\hat{f}(\chi)$ that can render $\hat{f}$ monotonic on the set $A \cup \{\chi\}$. Furthermore, we can identify a point $\chi_L \in L_\chi$ that minimises the distance to χ under the constraint that $\hat{f}(\chi_L) = y_L$. Likewise,

$$y_U = \min\left\{\hat{f}(x_{1i}, ..., x_{pi})\,;\, (x_{1i}, ..., x_{pi}) \in U_\chi\right\}$$

defines an upper limit for the permissible values of $\hat{f}(\chi)$, and we can select a point $\chi_U \in U_\chi$ that minimises the distance to χ under the constraint that $\hat{f}(\chi_U) = y_U$.

If χ is on the straight line between χ_L and χ_U, it would be natural to use linear interpolation to assign a value to $\hat{f}(\chi)$, in other words to set

$$\hat{f}(\chi) = y_L + \frac{\|\chi - \chi_L\|}{\|\chi_U - \chi_L\|}(y_U - y_L).$$

where $\|\boldsymbol{u}\|$ denotes the length of the vector $\boldsymbol{u}$. Regardless of the location of χ we can set

$$\hat{f}(\chi) = y_U + \frac{(\chi - \chi_L, \chi_U - \chi_L)}{(\chi_U - \chi_L, \chi_U - \chi_L)}(y_U - y_L)$$

where $(\boldsymbol{u},\boldsymbol{v})$ denotes the scalar product of the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$.

If L_x or U_x is empty, we assign values to $\hat{f}(\chi)$ as follows:

$$\hat{f}(\chi) = \begin{cases} y_U, \text{ if } L_\chi \text{ is empty and } U_\chi \text{ is nonempty} \\ y_L, \text{ if } U_\chi \text{ is empty and } L_\chi \text{ is nonempty} \\ \bar{y}, \text{ if both } L_\chi \text{ and } U_\chi \text{ are empty} \end{cases}$$

where $\bar{y}$ is the mean response for the elements in the entire estimation set or a subset of elements within a fixed distance to χ.

The procedure described above can be repeated for an arbitrary set of points. However, it is important to note that the estimation set A must be updated from A to $A \cup \{\chi\}$ each time $\hat{f}$ has been extrapolated to a new point χ. Otherwise, there may be pairs of extrapolated values for which the monotonicity is violated.

If the evaluation set is large, the above-mentioned procedure can be computationally cumbersome. Hence there is also a need for extrapolation procedures that can provide an approximately monotonic response surface over a large set of points. For example, it can be convenient to use kernel smoothing or locally weighted scatter-plot smoothing [17] to extrapolate the fitted responses in a monotonic regression to a fine grid of values of the explanatory variables (see Figure 1b).

4.3 Normalisation of contaminants in fish

Simple time series plots of observed concentrations of mercury in Atlantic cod caught in the middle of the North Sea (53° 10' N, 2° 5' E) indicate a downward trend (Figure 4a). However, this may, at least in part, be a spurious trend caused by temporal changes in the lengths of the analysed fish. Hence, it is of great interest to normalise the observed mercury concentrations with respect to fish length. Figure 4b illustrates the results obtained by using the normalisation procedure described in sections 4.1 and 4.2. Apparently, the mercury trend after normalisation is considerably smaller than in the raw data.

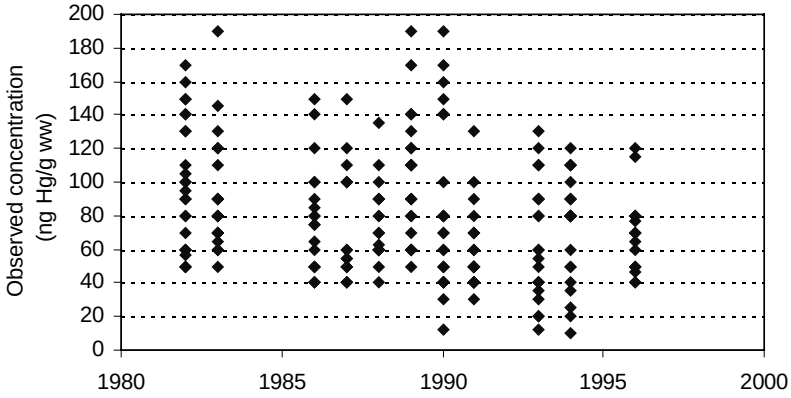


Figure 4a. Observed concentrations of mercury in muscle tissue from Atlantic cod (*Gadu morhua*) caught in the North Sea (53^o 10'N, 2^o 5'E).

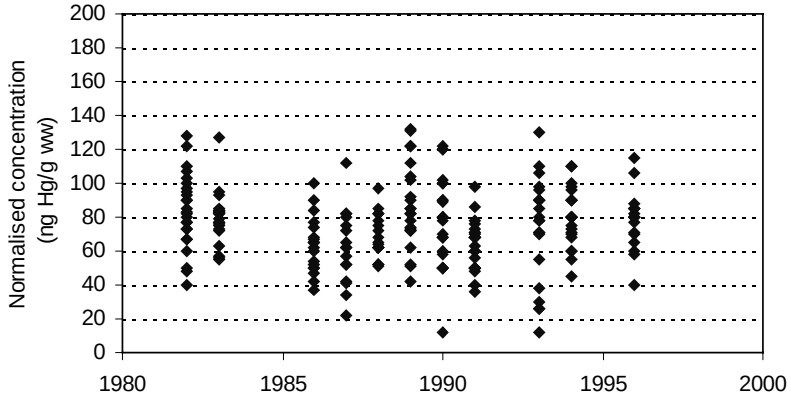


Figure 4b. Concentration of mercury in Atlantic cod (*Gadu morhua*) normalised to a body length of 49.6 cm.

5 DISCUSSION

In this article, we demonstrate that the recently developed GPAV algorithm for monotonic regression has a great potential for applications in environmental science and management. In many cases, the assumption of monotonic response has a solid foundation in process-based modelling or strong empirical evidence, whereas in other cases, such as trend assessment of environmental quality data, it can be more correct to regard monotonicity as a reasonable simplification. In either case, monotonic regression is convenient to use because issues concerning model selection are reduced to a minimum. Conventional parametric modelling of non-linear responses to two or more variables is usually more demanding, and the fit to observed data can be unsatisfactory even if a large number of models are tested. Kernel smoothing, locally weighted scatter-plot smoothing, and other non-parametric regression techniques are often more viable alternatives to monotonic regression. However, the latter

methods may produce very odd results if the set of points for which responses have been observed is very unevenly distributed. In addition, it may be unsatisfactory to obtain a non-monotonic response surface when there is theoretical support for monotonicity.

The calculations that were undertaken to simultaneously extract a monotonic trend and seasonal components from a time series of water quality data illustrate that monotonic regression usually produces a relatively smooth response surface. In this respect, our method is similar to the currently used procedures for time series decomposition (e.g., [15]). However, the estimated seasonal effects can be different, because, in contrast to other techniques, our method includes constraints on how such effects can vary from one season to the next.

The normalisation in section 4 represents a type of statistical analysis that has attracted increasing interest in environmental monitoring. Most of the methods presently in use are based on linear regression models for the removal of irrelevant variation in collected data [18], and model selection studies have shown that such models can perform well even if some of the underlying processes are non-linear [19]. However, it has also been demonstrated that non-linear features of a normalisation model can improve the performance if there is a strong, non-linear trend in the analysed data [20]. The monotonic regression described in this article provides yet another tool for normalisation that can be particularly useful if the natural fluctuations in collected data include non-linear responses to one or more covariates.

6 ACKNOWLEDGEMENTS

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

7 REFERENCES

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

[2] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical inference under order restrictions*. New York: Wiley, 1972.

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

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

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

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

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

[8] O. Burdakov, A. Grimvall, and M. Hussian. A generalised PAV algorithm for monotonic regression in several variables. *Proceedings of COMPSTAT*, 23-27 Aug, 2004, Prague, 2004a.

[9] O. Burdakov, O. Sysoev, A. Grimvall, and M. Hussian. An algorithm for isotonic regression problems. This Proceedings, 2004b.

[10] H. Mukarjee. Monotone nonparametric regression. *The Annals of Statistics*, **16**, 741-750, 1988.

[11] H. Mukarjee and H. Stem. Feasible nonparametric estimation of multiargument monotone functions. *Journal of the American Statistical Association*, **425**, 77-80, 1994.

[12] M. Strand. Comparison of methods for monotone nonparametric multiple regression. *Communications in Statistics - Simulation and Computation*, **32**, 165-178, 2003.

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

[14] D. Gamarnik. Efficient learning of monotone concepts via quadratic optimization. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 24-26 Jul 1998, Madison, Wisconsin, pp. 134-143, 1998.

[15] A. Maravall and V. Gómez. Estimation, prediction and interpolation for nonstationary series with the Kalman filter. *Journal of the American Statistical Association*, **89**, 611-624, 1994.

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

[17] W. Härdle. *Applied non-parametric regression*. Cambridge: Cambridge University Press, 1997.

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

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

[20] M. Hussian, A. Grimvall, and W. Petersen. Estimation of the human impact on nutrient loads carried by the Elbe River. Accepted for publication in *Journal of Environmental Monitoring and Assessment*, 2004.