2003

Longitudinal analysis of bioaccumulative contaminants in freshwater fishes

Jianguo Sun
University of Missouri

C. J. Schmitt
US Geological Survey, Columbia Environmental Research Center, cjschmitt@usgs.gov

Environmental and Ecological Statistics **10,** 419–428, 2003

Longitudinal analysis of bioaccumulative contaminants in freshwater fishes

JIANGUO SUN,[1] YANGJIN KIM[1] and CHRISTOPHER J. SCHMITT[2]

[1]*Department of Statistics, University of Missouri, 222 Math Sciences Building, Columbia, MO 65211 U.S.A.*
[2]*U.S. Geological Survey, Columbia Environmental Research Center, 4200 New Haven Road, Columbia, MO 65201 U.S.A.*

Received September 2001; Revised March 2003

The National Contaminant Biomonitoring Program (NCBP) was initiated in 1967 as a component of the National Pesticide Monitoring program. It consists of periodic collection of freshwater fish and other samples and the analysis of the concentrations of persistent environmental contaminants in these samples. For the analysis, the common approach has been to apply the mixed two-way ANOVA model to combined data. A main disadvantage of this method is that it cannot give a detailed temporal trend of the concentrations since the data are grouped. In this paper, we present an alternative approach that performs a longitudinal analysis of the information using random effects models. In the new approach, no grouping is needed and the data are treated as samples from continuous stochastic processes, which seems more appropriate than ANOVA for the problem.

Keywords: longitudinal data analysis, National Contaminant Biomonitoring Program, organochlorine pesticide, temporal trends

1. Introduction

The analysis of bioaccumulative contaminants in freshwater fish is an important component of environmental contaminant monitoring programs (Messer *et al.*, 1991; Goldstein *et al.*, 1996; Schmitt *et al.*, 1999). Among others, the National Contaminant Biomonitoring Program (NCBP) was initiated in 1967 as a component of the National Pesticide Monitoring program. It consists of periodic collection of freshwater fish and other samples and the analysis of the concentrations of persistent environmental contaminants in these samples. The NCBP provided information on the success of legislative and regulatory actions intended to reduce environmental concentrations of bioaccumulative toxins and on the effects of changing agricultural practices.

In the NCBP, fishes were collected from over 100 stations located at key points in major rivers throughout the United States and in the Great Lakes. From 1976 through 1981, the collections were made at about half the stations in the fall of even-numbered years and the rest in the fall of odd-numbered years. That is, one collection cycle required more than a

year for the completion. In 1984–1985, the samples were collected in fall and winter 1984 and early spring 1985 and the same was done for the 1986–1987 cycle. After 1986–1987 cycle, the NCBP was suspended and replaced by another program which covered a wider array of contaminants and included biological endpoints.

The NCBP freshwater fish data contain concentration information obtained from composite fish samples on many chemicals. For example, the concentrations of many organochlorine residues such as the insecticide toxaphene and polychlorinated biphenyls (PCB), which are industrial contaminants, were measured during the program period. It was found that the mean concentrations of total PCBs declined significantly nationally. In contrast, the mean concentrations of toxaphene in NCBP freshwater fishes were found to increase steadily through the mid-1970s, peaked around 1980, and declined significantly in each collection cycle thereafter (Schmitt *et al.*, 1990, 1999). In this paper, we will confine our analysis and discussion to 2,2-bis (p-chlorophenyl)-1,1-dichloroethylene (p,p'-DDE), which is a metabolite and a major component of the organochlorine pesticide DDT and will be referred to as DDE, with focus on the temporal trend of its concentrations during the program period. The methods discussed below also apply to other chemicals. For more details about the NCBP such as sample collection and laboratory measurement procedures, see Schmitt *et al.* (1981, 1990, 1999).

To analyze NCBP data and assess the trend of concentrations of environmental contaminants, a common approach has been to apply the mixed two-way ANOVA model to compare the means of concentrations over different measurement cycles or combined cycles (Schmitt *et al.*, 1990, 1999). This is approximately equivalent to using t-tests to compare every two adjacent cycles. More specifically, Schmitt *et al.* (1999) employed Fischer's protected LSD to test for concentration changes at the national level and planned contrasts and the Mann–Kendall test to test for trends at individual NCBP stations. They also used simple linear regression and analysis of covariance techniques for the analysis. One major disadvantage of these methods is that it ignores the timing of each concentration measurement by treating the concentration measurements within a cycle as being measured at the same measurement time. In other words, data grouping is used, which causes the loss of detailed information about the concentration process. By using cycles instead of real measurement times, the method misses the important feature of the data: the concentration is a continuous function of time. In particular, it does not give the temporal trend of the concentrations. Also the existing methods usually assume either no or constant serial correlation among contaminant concentrations from the same stations. Corresponding to these, an alternative approach is presented that overcomes the shortcomings of the existing methods and allows a longitudinal analysis of the data.

The main goal of this paper is to propose a simple approach for the analysis of NCBP that can capture the temporal trend of the contaminants in freshwater fish, to relax the assumptions required by existing methods and to make use of all available information. For this, we will treat NCBP fish concentration data as longitudinal data and present longitudinal data analysis techniques for the assessment of temporal trends of contaminant concentrations. We will begin in the next section with introducing notation and discussing the method for estimating the average overall temporal trend of the concentrations. For this, a kernel estimation procedure is presented. Section 3 presents two random effects models for studying the temporal trend at each individual station. One is a simple linear random effects model and the other is a nonparametric random effects model. The presented methods need no grouping and treat measured concentrations as realizations of

underlying longitudinal stochastic processes, which is more natural than the ANOVA model for the problem. Section 4 applies the presented techniques to the concentrations of DDE in the NCBP. The proposed methods give more insights than the mixed two-way ANOVA method and yield smooth estimates of concentrations in time that cannot be obtained using the existing approach. Section 5 contains conclusions and discussions.

2. Notation and estimation of overall temporal trend

Suppose that there are n stations from which freshwater fish samples are collected and concentrations of a contaminant in the samples are measured. Let $Y_i(t)$ be a stochastic process denoting the log concentration of the contaminant at time t for the ith station, $i = 1, \ldots, n$. Suppose that observed data have the form $\{y_{ij} = Y_i(t_{ij});\ j = 1, \ldots, m_i,\ i = 1, \ldots, n\}$, where $t_{i1} < \cdots < t_{im_i}$ denote the measurement times of the concentrations for station i. For the NCBP data, the time t is years. In the following analysis, we are mainly interested in two problems. One is to estimate the overall average temporal trend of the concentration of the contaminant under study and the other is to study the concentration trend for each individual station.

In this section, we will consider estimation of the overall temporal trend of concentrations of a contaminant. For this, assume that the Y_i's come from a homogeneous population and let $\mu(t) = E[Y_i(t)]$ denote the mean function of the concentration process, which is assumed to be an arbitrary smooth function of time. To estimate $\mu(t)$, we will adopt the locally adaptive kernel estimation method (Hart and Wehrly, 1986; Muller and Stadtmuller, 1987). Let $K(u)$ be a nonnegative symmetric function around $u = 0$ with $\int_{-\infty}^{\infty} K(t)\,dt = 1$, which is usually called a kernel function. Also let h be a positive parameter called the bandwidth parameter, which determines how large a neighborhood of t is used to calculate the local average. Define

$$w_{ij}^*(t, h) = h^{-1} K\left\{\frac{(t_{ij} - t)}{h}\right\},$$

and

$$w_{ij}(t) = \frac{w_{ij}^*(t, h)}{\sum_{u=1}^{n} \sum_{l=1}^{m_u} w_{ul}^*(t, h)},$$

$j = 1, \ldots, m_i,\ i = 1, \ldots, n$. Then the kernel estimate of $\mu(t)$ is given by

$$\hat{\mu}(t) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}(t) y_{ij}. \tag{1}$$

Many kernel functions can be used. In the following, we will use the Gaussian kernel

$$K(t) = \exp(-u^2/2).$$

For a given kernel function K, one needs to choose the bandwidth parameter h. One common way for this, which is used below, is to choose h such that it minimizes

$$SSR(h) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{y_{ij} - \hat{\mu}(t_{ij})\}^2 \tag{2}$$

(Rice and Silverman, 1991).

We remark that the above kernel estimate of the overall temporal trend is similar to the mixed ANOVA estimate in that both are based on averages. However, there exist significant differences. One is that the former uses weighted averages and is more flexible in terms of the way how the average is taken, while the latter uses simple empirical averages based on observations within the same cycle. One problem for the average based on cycles is that the measurement times may be quite different from station to station within a cycle and a measurement time at one station could be closer to measurement times at other stations in different cycles than the measurement times at other stations within the same cycle. It is apparent that in this case, the ANOVA method would give biased estimate. Another difference is that the kernel method gives a smooth estimate as it should be since the underlying concentration is continuous in time, while the ANOVA method can only provide a rough direction. Also the ANOVA approach treats concentration measurements at different times within the same station as irrelevant and estimates the temporal trend at different time points independently, which is against the relevant and continuous nature of concentrations.

3. Estimation of concentration trends for individual stations

In this section, we will investigate the temporal trend of concentrations of a contaminant for individual stations. For this purpose, we will consider two random effects models. One is the simple linear random effects model:

$$Y_i(t_{ij}) = \beta_0 + \beta_1(t_{ij} - 1975) + b_{0i} + b_{1i}(t_{ij} - 1975) + e_{ij}, \tag{3}$$

$j = 1, \ldots, m_i$, $i = 1, \ldots, n$. In the above, $\beta = (\beta_0, \beta_1)'$ are fixed parameters, $b_i = (b_{0i}, b_{1i})'$ represent random effects, and $e_i = (e_{i1}, e_{i2}, \ldots, e_{im_i})'$ are random errors. In the model, the 1975 (year) is used because the first year at which concentration data are available is 1976.

The above model assumes that the concentration is on average a linear function of time and differences among concentrations from different stations can be represented by random effects which are also linear functions of time. It is worth noting that model (3) is similar to, but different from, that used in the mixed two-way ANOVA analysis. The key difference is that the model used in the ANOVA analysis has no time effect. Also note that although there is probably no concentration strictly following a linear model, model (3) provides a simple and good approximation to the true temporal trend.

Let $Y_i^* = \{Y_i(t_{i1}), \ldots, Y_i(t_{im_i})\}'$ and

$$Z_i = \begin{pmatrix} 1 & t_{i1} - 1975 \\ . & . \\ . & . \\ . & . \\ 1 & t_{im_i} - 1975 \end{pmatrix}.$$

Then model (3) can be rewritten as

$$Y_i^* = Z_i\beta + Z_i b_i + e_i,$$

$i = 1, \ldots, n$. Note that this model was first proposed by Laird and Ware (1982) and is commonly used in medical fields due to its simplicity.

The above model can be generalized to

$$Y_i(t_{ij}) = \mu(t_{ij}) + b_{0i} + b_{1i}(t_{ij} - 1975) + e_{ij}, \tag{4}$$

$j = 1, \ldots, m_i,\ i = 1, \ldots, n$, where $\mu(t)$ is the mean function defined in the previous section. Let $\boldsymbol{t}^0 = (t_1^0, \ldots, t_r^0)$ be the vector of ordered distinct values of the time points $\{t_{ij}, j = 1, \ldots, m_i,\ i = 1, \ldots, n\}$ and $\theta_\mu = (\mu(t_1^0), \ldots, \mu(t_r^0))'$. Also let W_i be a $m_i \times r$ indicator matrix whose (j,l)th element is 1 if $t_{ij} = t_l^0$ and 0 otherwise, $j = 1, \ldots, m_i,\ l = 1, \ldots, r$. Then model (4) can be rewritten as

$$Y_i^* = W_i\theta_\mu + Z_i b_i + e_i, \quad i = 1, \ldots, n.$$

As mentioned before, model (3) approximates the concentration of a contaminant using a linear function of time t. In contrast, model (4) assumes that the concentration is a smooth nonlinear function of time t. Both models suppose that the trends of concentrations across all stations have a common baseline mean function and their difference can be characterized by random effects b_i's.

To make inference about the parameters in models (3) and (4), as usual, we will assume that the b_i's and e_i's are mutually independent and

$$b_i \sim N(0, \Sigma), \quad e_i \sim N(0, \sigma^2 I_i),$$

$i = 1, \ldots, n$, where Σ is a 2 by 2 matrix, σ^2 is an unknown parameter and I_i denotes the $m_i \times m_i$ identical matrix. Let $p = (\beta, \Sigma, \sigma^2)$ for model (3) and $p = (\theta_\mu, \Sigma, \sigma^2)$ for model (4). Under the model (3), the parameters p can be easily estimated by the maximum likelihood method or the restricted maximum likelihood method under the above assumptions and the estimation procedure is available in many statistical softwares. For the analysis here, the procedure in SAS is used. Under the model (4), to estimate p, we propose first to estimate θ_μ using the method given in the previous section and then to apply the method for model (3) to the modified data set $\{y_{ij}^* = Y_i(t_{ij}) - \hat{\mu}(t_{ij}),\ j = 1, \ldots, m_i,\ i = 1, \ldots, n\}$.

Once the estimates, $\hat{p}$, of p are obtained, the random effects parameters b_i's can be estimated by its conditional expectation $\hat{b_i} = E(b_i | Y_i^*, \hat{p})$. It follows that the estimated temporal trend of the concentration of a contaminant for the ith station is given by

$$\hat{Y}_i(t_l^0) = \hat{\beta}_0 + \hat{\beta}_1(t_1^0 - 1975) + \hat{b}_{0i} + \hat{b}_{1i}(t_l^0 - 1975), \tag{5}$$

based on model (3), or

$$\hat{Y}_i(t_l^0) = \hat{\mu}(t_l^0) + \hat{b}_{0i} + \hat{b}_{1i}\ (t_l^0 - 1975), \tag{6}$$

based on model (4), $l = 1, \ldots, r$.

4. Analysis of concentrations of DDE

Now we apply the techniques presented in the previous sections to analyzing concentrations of DDE collected from NCBP from 1976 to 1987. In the NCBP, the

samples are generally composites of five whole, adult fishes of the same species and the concentrations used in the analysis are unweighted means of the concentrations of several samples from the same stations. Readers are referred to Schmitt *et al*. (1999) for more details on the program and the data set, in particular the procedures used to collect fish samples and to measure concentrations and the information about stations. The analysis here will be based on 548 mean concentration measurements available from 113 stations in major rivers in the United States and in the Great Lakes. Following Schmitt *et al*. (1999) and others, we will analyze the log-transformed concentrations of DDE. The raw NCBP data can be obtained at http://www.crec.usgs.gov/data/data.htm.

To give an idea about the pattern of observed concentrations, Fig. 1 presents log DDE observed from 5 randomly selected stations with measurements from the same station connected by straight (dotted or broken) lines. Also included in Fig. 1 is the kernel estimate $\hat{\mu}(t)$ (solid line) of the overall temporal trend of DDE concentrations given by (1). For the data set here, the bandwidth parameter $h = 1.05$ was selected using the criterion (2) and used. It can be seen from Fig. 1 that on average the DDE concentration was stable from 1976 to 1979 and then decreased. Note that the estimate suggests that the DDE concentration was going up from 1986 to 1987. However, caution should be used here since only 8 DDE measurements (stations) are available in 1987.

To study the DDE concentration for individual stations, we first fitted the observed data to model (3) and obtained $\hat{\beta}_0 = -1.7855$ and $\hat{\beta}_1 = -0.0804$ with the estimated standard deviations being 0.1050 and 0.0065, respectively. This again suggests that the DDE concentration was decreasing during the period 1976–1987. For most stations, the estimated random intercept effects (b_{0i}) are significantly different from zero under the significance level of 0.05 and this indicates that the DDE concentrations among stations in 1975 were quite different from each other. On the other hand, the estimated random slope effects (b_{1i}) are significantly different from zero for only a few stations, which are stations

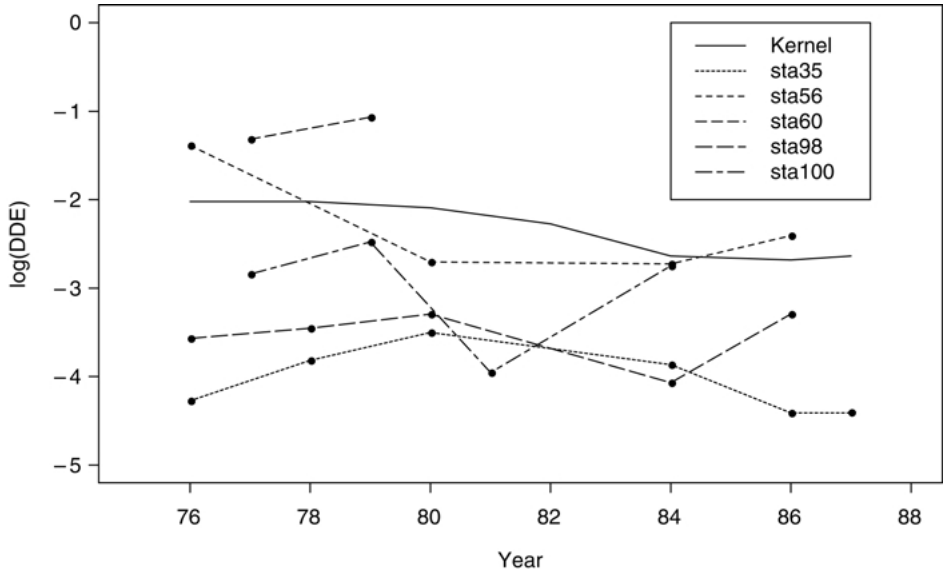

Figure 1. Observed concentrations of log(DDE) and Kernel estimate.

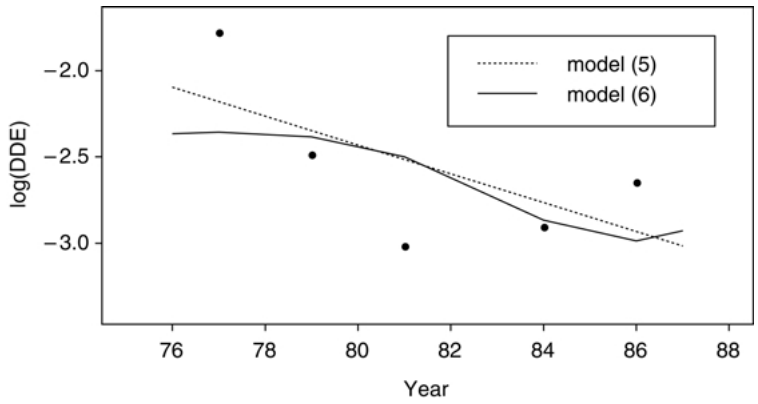

Figure 2. Observed and prediced log(DDE) for station 15.

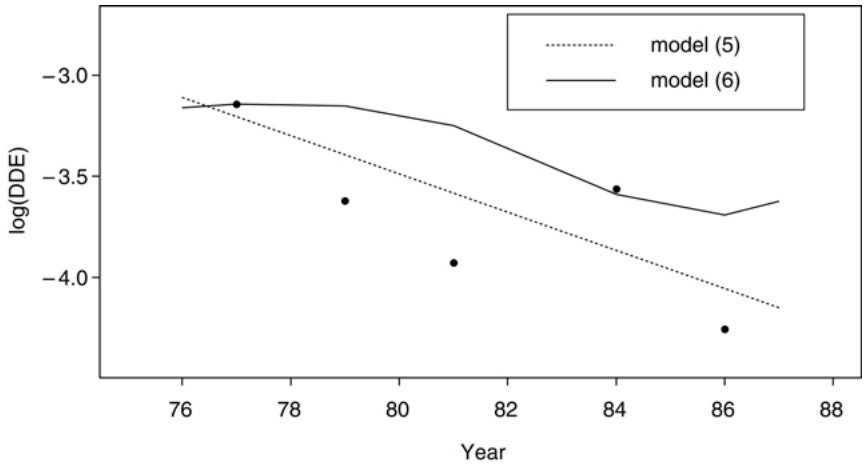

Figure 3. Observed and prediced log(DDE) for station 85.

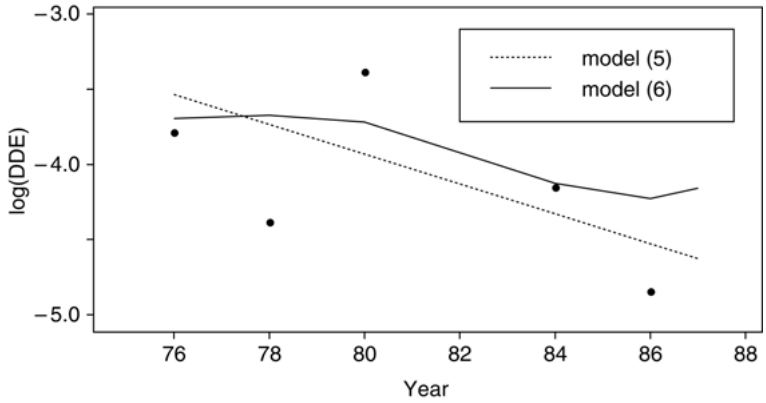

Figure 4. Observed and prediced log(DDE) for station 93.

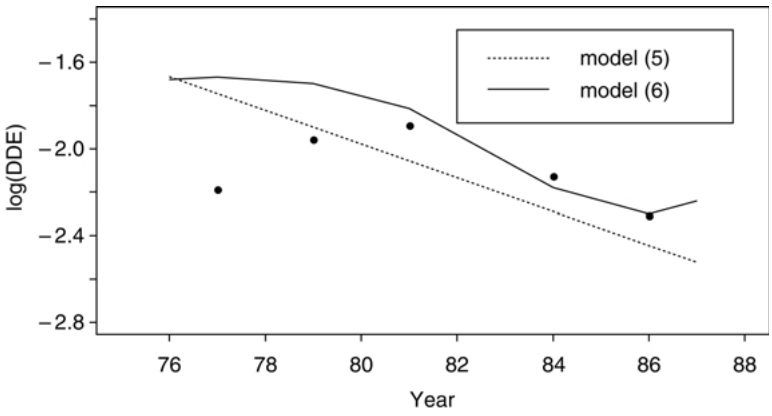

Figure 5. Observed and prediced log(DDE) for station 108.

14, 16, 21, 32, 47, 51, 74, 80, 84, 93, 104, 105, 115, and 116. To see graphically the temporal trends of the DDE concentration for individual stations, Figs 2–5 display the estimated trends (dotted lines) given by (5) along with observed DDE concentrations for stations 15, 85, 93, and 108, respectively. It can be seen that for all four stations, the DDE concentrations were decreasing during the period. In fact, this is true for all stations since all of the absolute values of the estimated random slopes are smaller than the absolute value of $\hat{\beta}_2$.

We then fitted observed data to model (4). In this case, none of the estimated random intercept effects (b_{0i}) is significantly different from zero at the significance level of 0.05. There are 18 stations for which the estimated slope random effects (b_{1i}) are significantly different from zero. They are stations 16, 21, 28, 32, 35, 47, 50, 51, 71, 74, 77, 80, 84, 93, 104, 105, 115, and 116. To compare the estimated temporal trends of the DDE concentration given by the models (3) and (4), the estimated temporal trends (solid lines) of the DDE concentration given by (6) are presented in Figs 2–5. As model (3), the estimated trends under the model (4) also suggest that the DDE concentrations were steadily decreasing for the observation period.

5. Conclusions and discussions

This paper considered statistical analysis of concentrations of bioaccumulative environmental contaminants with focus on the DDE concentrations measured in NCBP during the period 1976–1987. For the analysis, we presented parametric, semiparametric and nonparametric methods that take into account the longitudinal nature of the concentrations, which was not the case for existing methods. Also the presented methods are more natural and give more insights than the existing approach. In particular, the methodology yields estimates of the temporal trends of concentrations of chemical contaminants. In contrast, the existing ANOVA method does not really provide an estimate of the temporal trend, but only gives independent averages of concentrations within each cycle. If the measurement times within a cycle are close to each others and their relative locations from station to station stay roughly the same from cycle to cycle, then the

averages given by ANOVA could provide rough direction of concentration changes from cycle to cycle. As mentioned before, however, this is not quite the case for NCBP and thus the ANOVA average could give biased or misleading directions for concentration changes as pointed below.

The analysis of DDE concentrations suggests that although observed DDE concentrations were up and down for some stations, they were decreasing during the period 1976–1987 both on the overall average and for each individual stations as expected. In comparison, the results obtained using the existing approach suggest that the DDE concentration was decreasing for a few stations, increasing for one station, and not significantly changing for all other stations (Schmitt *et al*., 1999). We believe that the major reason for the difference here is that the presented approach can better make use of available information than the existing approach, which mainly bases the analysis on summary statistics.

It should be noted that as most of environmental data, chemical concentration data from NCBP have both serial and spatial correlations. As the existing approach, the inference procedures given in Section 3 requires that concentration measurements from different stations are independent of each other. We realized that this could introduce certain problems. However, we believe that if the main focus is on the mean or the temporal trend of the concentrations as it is here, the impact of ignoring spatial correlation is minor and especially, the estimates obtained in this way should still be unbiased. As mentioned before, the major goal here is to develop methods that can better deal with the serial correlation in NCBP data than the existing methods. As future research, one could consider two-dimensional longitudinal analysis.

In the above, we considered a simple linear model and a nonparametric model. Some other models could also be considered for analyzing concentration data. For example, one could fit the data to the linear model with some change points. This may be useful given that the concentration could be a linear function of time, but may change directions at suggested by the curve given in Fig. 1. Another alternative is to add extra terms to models (3) and (4) such as covariate effects when there exist covariates. Also the linear random effects in models (3) and (4) could be replaced by some nonlinear random effects. One difficulty that could occur in analyzing chemical contaminant data is the existence of right- or interval-censoring (Sun, 1998). For example, this is often the case for the measurements of PCB concentrations. However, no censoring was involved for concentration measurements of DDE, a virtually ubiquitous globe pollutant.

Acknowledgments

The authors wish to thank the editor and referees for their helpful comments and criticisms that improved the paper.

References

Goldstein, R.M., Brigham, M.E., and Stauffer, J.C. (1996) Comparison of mercury concentrations in liver, muscle, whole bodies, and composites of fish from the Red River of the North. *Can. J. Fish Aquat. Sci*., **53**, 244–52.

Hart, J.D. and Wehrly, T.E. (1986) Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.*, **81**, 1080–8.

Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–74.

Messer, J.J., Linthurst, R.A., and Overton, W.S. (1991) An EPA program for monitoring ecological status and trends. *Environ. Monit. Assess.*, **17**, 67–78.

Muller, H.G. and Stadtmuller, U. (1987) Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610–25.

Rice, J.A. and Silverman, B.W. (1991) Estimating the mean and covariate structure nonparametrically when the data are curves. *J. R. Statist. Soc. B*, **53**, 233–43.

Schmitt, C.J., Ludke, J.L., and Walsh, D. (1981) Organochlorine residues in fish, 1970–1974: National Pesticide Monitoring Program. *Pestic. Monit. J.*, **14**, 136–206.

Schmitt, C.J., Zajicek, J.L., May, T.W., and Cowman, D.F. (1999) Organochlorine residues and elemental contaminants in U.S. freshwater fish, 1976–1986: National Contaminant Biomonitoring Program. *Rev. Environ. Contam. Toxicol.*, **162**, 43–104.

Schmitt, C.J., Zajicek, J.L., and Peterman, P.L. (1990) National Contaminant Biomonitoring Program: residues of organochlorine chemicals in freshwater fishes of the United States, 1976–1984. *Arch. Environ. Contam. Toxicol.*, **19**, 748–82.

Sun, J. (1998) Interval censoring. *Encyclopedia of Biostatistics*, Wiley, 2090–5.

Biographical sketch

Jianguo Sun is an Associate Professor and Yangjin Kim is a Ph.D. candidate at the Department of Statistics, University of Missouri. Christopher Schmitt is a Research Fishery Biologist at the Columbia Environmental Research Center of U.S. Geological Survey. Dr Sun's research interests include survival analysis and longitudinal data analysis and he is interested in applying statistical techniques in these fields to environmental problems such as one discussed in the paper. Dr Schmitt has been working at the Columbia Environmental Research Center more than 20 years and published a number of papers in referred journals on evaluating National Contaminant Biomonitoring Program and analyzing data about chemical contaminants in freshwater fishes.