Smoothing Parameter Selection Methods for Nonparametric Regression with Spatially Correlated Errors

M. Francisco-Fernández

Departamento de Matemáticas, Facultad de Informática

Universidad de A Coruña, La Coruña,15071, Spain.

and

J. D. Opsomer[*]

Department of Statistics, Iowa State University, Ames, IA 50011, USA.

October 14, 2003

Summary. Nonparametric regression makes it possible to visualize and describe spatial trends without requiring the specification of a parametric model, but appropriate choice of smoothing parameters is important to avoid misinterpreting the nonparametric fits. Because spatial data are often correlated, currently available data-driven smoothing parameter selection methods often fail to provide useful results. We propose to adjust the generalized cross-validation (GCV) criterion for the effect of the spatial correlation, and develop an approach to do so in the case of bivariate local polynomial regression. The adjustment uses a pilot fit to the data and the estimation of a parametric covariance model. The method is easy to implement, and we show that it leads to improved smoothing parameter selection results, even when the covariance

[*] *email:* jopsomer@iastate.edu

model is misspecified. The method is illustrated using water chemistry data collected in a survey of lakes in the Northeastern U.S.

Key words: local polynomial regression; generalized cross-validation (GCV), semivariogram estimation, Northeastern Lakes survey, acid neutralizing capacity (ANC).

1. Introduction

Many datasets in environmental and biological sciences contain observations that have spatial locations. Whether the locations are of scientific interest or not, it is often useful to perform an exploratory analysis by looking at the spatial distribution of study variables. Nonparametric regression methods are an appropriate class of tools to perform such exploratory analysis, because they do not require a specific parametric shape to be selected before fitting the data.

In this article, we consider the problem of visualizing the spatial pattern of *Acid Neutralizing Capacity* (ANC) for a sample of 338 lakes in the Northeastern U.S. ANC, also called *acid binding capacity* or *total alkalinity*, measures the buffering capacity of water against negative changes in pH (Wetzel, 1975, p. 172), and is often used as an indicator of the acidification risk of water bodies. Surface waters with ANC values below 200 μeq/L (as measured using the Gran titration method) are considered at risk of acidification (NAPAP 1991, p. 15). A number of factors determine the ANC value of surface water, including the characteristics of the soils and mineral formations of the watershed and the presence of acid producing sources such as bogs. However, it is generally believed that human-induced acidic deposition, in both its wet ("acid rain") and dry form, is a significant contributor to lake and stream acidification, and hence causes a decrease in ANC over time.

Government agencies have conducted surveys and monitoring studies to assess the effect of acidic deposition on surface waters, with the measurement of ANC one of the key variables (NAPAP 1991, Stoddard et al., 2003). The complex interactions between the natural and human-induced factors affecting ANC make it difficult to write down a model for ANC distribution over a large geographic region. In the absence of such a model, representing the distribution of ANC as a smooth spatial function provides a useful “first step” in identifying areas of high and low ANC. Such a spatial distribution map can be useful in the selection of monitoring locations, for instance, or in the development of more complete models incorporating substantive covariates.

In the current article, we will develop a spatial distribution map for ANC data collected between 1991 and 1996 during a survey of the ecological condition of lakes in the Northeastern states of the U.S. (see Figure 1), as part of the U.S. Environmental Protection Agency’s Environmental Monitoring and Assessment Program (Messer et al., 1991). We refer to Larsen et al. (2001) for a more detailed description of the survey design.

[Figure 1 about here.]

In order to fit a nonparametric regression model to spatial data such as the ANC lake measurements, values need to be specified for a set of smoothing parameters. In the case of local polynomial regression on data located over a geographic area, three smoothing parameters need to be selected. It is of course possible to leave these parameters at their default values specified by a software program, but there is no guarantee that this choice is even close to reasonable, and it might lead to a fit that masks important features in the data or one that exhibits spurious patterns. Conversely, the analyst might use “trial and error” and find satisfactory parameter values in this fashion. This approach is inefficient and prone to error, however.

It would clearly be preferable to be able to use a data-driven smoothing parameter selection method. Ideally, such a method would be easy to calculate and would select values for these parameters that provide a good trade-off between model smoothness and closeness of the fit to the data. In the spatial case considered here, the higher dimension of the model, combined with the likely presence of spatial correlation, makes data-driven smoothing parameter selection a difficult practical problem. Correlation in the spatial smoothing context can be thought of in two ways: either as "true" correlation caused by observations located close to each other being influenced by each other, or as a "placeholder effect" for the fact that the spatial mean model does not capture the influence of omitted covariates (of course, a combination of both situations is possible as well). Especially in the second case, the correlation is a nuisance effect, to be removed as much as possible from the representation of the spatial mean trend.

Unfortunately, most smoothing parameter selection methods do not perform well in the presence of correlated errors, as research in the one-dimensional case has shown (see Hart (1996) and Opsomer et al. (2001) for overviews). This article will discuss spatial model fitting with local linear regression and propose a smoothing parameter selection method based on the Generalized Cross-Validation (GCV) criterion (Craven and Wahba, 1979), suitably adjusted for the presence of spatial correlation. As will be clear after reading the article, the approach described here is applicable to other smoothing techniques such as regression and smoothing splines, to several other data-driven optimality criteria such as cross-validation (CV) and Akaike's Information Criterion (AIC), and to higher dimensional problems.

The proposed smoothing parameter selection approach can be summarized as follows. First, identify the bias of the data-driven criterion due to the correlation (this bias will be a function of the distribution of the errors). Next, fit a simple parametric model to residuals from a pilot fit to the data, and use this model fit to correct the bias of the criterion function. Finally, optimize the bias-corrected criterion to select smoothing parameter values. While this might not result in an "optimal" choice for the smoothing parameters, we will argue that this rather simple approach works well in practice, even if the parametric form of the correlation function is misspecified. It is certainly less cumbersome than a trial-and-error search, and works significantly better than using a criterion that does not correct for correlation.

[Figure 2 about here.]

[Figure 3 about here.]

To illustrate both the problems associated with using an unadjusted GCV criterion and the proposed solution, consider the local linear regression fits in Figures 2 and 3 (details on the estimation procedures and the fits will be discussed below). The first figure was obtained by applying a search algorithm to minimize the GCV criterion. The plot shows significant amount of variation that appears to be spurious, especially in the fit for the Western half of the data. It should be noted that the amount of variability shown here is actually less than that actually obtained by the regression, since the plot trims off several large “spikes” that go well beyond the values on the scale, both in the positive and negative direction (this was done to maintain comparability with Figure 3). In addition, the GCV algorithm had significant problems converging to a solution, since values very close to the GCV solution resulted in singular design matrices. In contrast, Figure 3 displays the fit obtained with the bandwidth that minimizes the correlation-adjusted GCV criterion. In this fit, singularity problems were avoided, so that the algorithm converged to a solution using fewer iterations. Compared to Figure 2, this fit is much smoother, but it still exhibits a pattern of higher mean ANC values in the Western portion of the plot and the regions of lower values in the North and East. We will discuss the fitted models in more detail in Section 5.

The idea of adjusting the data-driven optimality criterion for bandwidth selection in the presence of correlation is not new. Altman (1990) discussed this approach for the time series case and referred to it as the “direct method” of bandwidth selection under correlation (the “indirect method” consists of transforming the residuals and then applying the unadjusted criterion to the transformed residuals). Related approaches are those of Chiu (1989) and Hart (1991). For references to other approaches for bandwidth selection in the time series case, see Opsomer et al. (2001). The situation considered in this article is conceptually similar to that studied by these other authors. However, spatial data are often not regularly spaced as in time series, and different methods are needed to estimate the correlation function. Therefore, it is still of interest to study spatial smoothing separately from time series smoothing, and to develop a simple and practical bandwidth selection method for spatial smoothing. The method we propose is applicable to data that are either regularly or irregularly spaced, with the latter case illustrated by the Northeastern Lakes survey (see Figure 1).

As discussed above, the proposed bandwidth selection method requires estimation of the correlation function. We do this by calculating the empirical semivariogram of the residuals of a pilot fit (see Cressie, 1993, p.70), and using a method-of-moments estimator for the parameters of the correlation function. In this context, we show that the binned empirical semivariogram can be viewed as a kernel regression estimator of the true semivariogram function, and prove its consistency under a version of increasing-domain asymptotics. This result is of some theoretical interest separately from the bandwidth selection context, since we did not find a previously published proof for the consistency of the binned semivariogram estimator.

The outline of the remainder of the paper is as follows. In Section 2, we describe the statistical model and the nonparametric estimator, and we explain in more detail the smoothing parameter selection method. Section 3 discusses estimation of the semivariogram and the parameters of the correlation function. In Section 4, simulation experiments are used to evaluate the practical properties of our approach, and Section 5 discusses the application of the proposed method to the Northeastern Lakes data.

2. Smoothing Parameter Selection with Correlation-Adjusted GCV

2.1 *Review of Local Linear Regression for Spatial Data*

Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ be a set of $\mathbb{R}^{D+1}$-valued random vectors, where the Y_i are scalar responses variables and the $\boldsymbol{X}_i$ are $\mathbb{R}^D$-valued predictor variables with a common density f_x with compact support $\Omega \subseteq \mathbb{R}^D$. The multivariate nonparametric regression problem is that of estimating $m(x) = E\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right)$ at a location $\boldsymbol{x} \in \Omega$, where $m(\cdot)$ is not restricted to belong to a specific parametric family of functions. In this article, we assume the model

$$Y_i = m\left(\boldsymbol{X}_i\right) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\mathrm{E}(\varepsilon_i | \boldsymbol{X}_i) = 0, \mathrm{Var}\left(\varepsilon_i | \boldsymbol{X}_i\right) = \sigma^2, \mathrm{Cov}\left(\varepsilon_i, \varepsilon_j | \boldsymbol{X}_i, \boldsymbol{X}_j\right) = \sigma^2 \rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j)$ with $\rho_n(\boldsymbol{x})$ continuous, satisfying $\rho_n(\mathbf{0}) = 1$, $\rho_n(\boldsymbol{x}) = \rho_n(-\boldsymbol{x})$, and $|\rho_n(\boldsymbol{x})| \leq 1$, $\forall \boldsymbol{x}$. The subscript n in $\rho_n(\cdot)$ indicates that the correlation function has to "shrink" as the sample size $n \to \infty$. This will be made more precise below.

The local linear estimator for $m(\cdot)$ at $\boldsymbol{x}$ is the solution for α to the least squares minimization

$$\min_{\alpha, \boldsymbol{\beta}} \sum_{i=1}^{n} \left\{Y_i - \alpha - \boldsymbol{\beta}^T (\boldsymbol{X}_i - \boldsymbol{x})\right\}^2 K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}),$$

where $\boldsymbol{H}$ is a $D \times D$ symmetric positive definite matrix; K is a D-variate kernel and $K_{\boldsymbol{H}}(\boldsymbol{u}) = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}\boldsymbol{u})$. The bandwidth matrix $\boldsymbol{H}$ controls the shape and the size of the local neighborhood used for estimating $m(\boldsymbol{x})$. The local linear estimator can be written explicitly as

$$\widehat{m}(\boldsymbol{x}; \boldsymbol{H}) = e_1^T \left(\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y} \equiv \boldsymbol{s}_{\boldsymbol{x}}^T \boldsymbol{Y} \tag{2}$$

where e_1 is a length $(D+1)$ vector with 1 in the first entry and all other entries 0, $\boldsymbol{W}_x = \mathrm{diag}\left\{K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{x}), \ldots, K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{x})\right\}$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, and $\boldsymbol{X}_x$ is a matrix with ith row equal to $(1, (\boldsymbol{X}_i - \boldsymbol{x})^T)$.

For the case of uncorrelated data with a random design, Ruppert and Wand (1994) derived the asymptotic mean squared error (AMSE) formula for the multivariate local linear estimator. Opsomer et al. (2001) provide the corresponding results for $D = 2$ when the errors are correlated. Liu (2001) generalized those results to arbitrary D, and we briefly summarize those results below.

We introduce some notation. For any matrix $\boldsymbol{B}$, we use $\lambda_{\max}(\boldsymbol{B})$ and $\lambda_{\min}(\boldsymbol{B})$ to denote its maximum eigenvalue and minimum egeinvalue, respectively. Let $\mathcal{H}_g(\boldsymbol{u})$ represent the

Hessian matrix of a D-variate function $g(\cdot)$ evaluated at $\boldsymbol{u}$. Let $\mathcal{X}$ represent the sequence $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \ldots\}$. As commonly done in kernel regression, we will state all our results conditionally on $\mathcal{X}$ (see Ruppert and Wand (1994) for a discussion of this approach).

We require the following assumptions:

- *A* 1. The function $m(\cdot)$ is second order differentiable on a compact set Ω, the error variance $\sigma^2 > 0$, and the design of the locations $f_x(\cdot) > 0$ on Ω.

- *A* 2. $K(\cdot)$ is symmetric, Lipschitz continuous and $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$, $\int \boldsymbol{u}K(\boldsymbol{u})d\boldsymbol{u} = \mathbf{0}$ and $\int \boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u})d\boldsymbol{u} = \mu_2(K)\boldsymbol{I}$ with $\mu_2(K) \neq 0$.

- *A* 3. The bandwidth matrix $\boldsymbol{H}$ is symmetric and positive definite. As $n \to \infty$, $\boldsymbol{H} \to 0$. The ratio $\lambda_{\max}(\boldsymbol{H})/\lambda_{\min}(\boldsymbol{H})$ is bounded above, and $n\,|\boldsymbol{H}|\,\lambda^2_{\min}(\boldsymbol{H}) \to \infty$ as $n \to \infty$.

- *A* 4. For the correlation function $\rho_n(\cdot)$, there exist constants ρ_I, C_ρ such that $n \int \rho_n(\boldsymbol{u})d\boldsymbol{u} \to \rho_I$ and $n \int |\rho_n(\boldsymbol{u})|\, d\boldsymbol{u} \leq C_\rho$. For any sequence $\varepsilon_n > 0$ satisfying $n^{1/D}\varepsilon_n \to \infty$,

$$n \int_{\|\boldsymbol{u}\| \geq \varepsilon_n} |\rho_n(\boldsymbol{u})|\, d\boldsymbol{u} \to 0 \quad \text{as } n \to \infty.$$

Assumption A4 implies that the integral of $|\rho_n(\boldsymbol{x})|$ should vanish as $n \to \infty$, and the vanishing speed should not be slower than $O\left(1/n\right)$. A4 further implies that the integral of $|\rho_n(\boldsymbol{x})|$ is essentially dominated by the values of $\rho_n(\boldsymbol{x})$ near to the origin $\mathbf{0}$. Hence, the correlation is short-range and decreases as $n \to \infty$. Arguing somewhat loosely, this can be considered a case of increasing-domain spatial asymptotics (Cressie, 1993, p.100), since this setup can readily be transformed to one in which the correlation function ρ_n is fixed with respect to the sample size, but the support Ω for $\boldsymbol{x}$ expands. The current setup with fixed domain Ω and shrinking ρ_n is more natural to consider when the primary purpose of the estimation is a fixed mean function $m(\cdot)$ defined over a spatial domain, not the correlation function itself.

Two examples of commonly used correlation functions that satisfy the conditions of assumption A4 for any $D \geq 1$ are the *exponential model*

$$\rho_n(\boldsymbol{x}) = \exp\left(-\alpha n \left\|x\right\|\right) \tag{3}$$

and the *rational quadratic model*

$$\rho_n(\boldsymbol{x}) = \frac{1}{1+\alpha\left(n\left\|x\right\|\right)^2}\ ,$$

with $\alpha > 0$ in both cases (Cressie, 1993, p.61).

Under assumptions A1–A4 and for $\boldsymbol{x}$ an interior point in Ω, Liu (2001) showed that

$$\mathrm{E}\left\{\widehat{m}(\boldsymbol{x};\boldsymbol{H}) - m(\boldsymbol{x})\middle|\,\mathcal{X}\right\} = \frac{1}{2}\mu_2(K)\mathrm{tr}\left(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{x})\right) + o_p\left(\mathrm{tr}(\boldsymbol{H}^2)\right) \tag{4}$$

$$\mathrm{Var}\left\{\widehat{m}(\boldsymbol{x};\boldsymbol{H})\middle|\,\mathcal{X}\right\} = \frac{\mu\left(K^2\right)\sigma^2(1+f(\boldsymbol{x})\rho_I)}{n\left|\boldsymbol{H}\right|f(\boldsymbol{x})} + o_p\left(\frac{1}{n\left|\boldsymbol{H}\right|}\right), \tag{5}$$

generalizing the results of Ruppert and Wand (1994) to the correlated error case. Defining AMSE($\widehat{m}(\boldsymbol{x};\boldsymbol{H})$) as the mean squared error approximation obtained by combining the leading terms of (4) and (5), the asymptotically optimal local bandwidth matrix can be computed by minimizing AMSE($\widehat{m}(\boldsymbol{x};\boldsymbol{H})$) with respect to $\boldsymbol{H}$. Liu (2001) showed that this minimizer is

$$\boldsymbol{H}_{opt}(\boldsymbol{x}) = \left\{\frac{\mu\left(K^2\right)\sigma^2(1+f(\boldsymbol{x})\rho_I)\left|\widetilde{\mathcal{H}}_m(\boldsymbol{x})\right|^{1/2}}{n\ D\ \mu_2^2(K)f(\boldsymbol{x})}\right\}^{1/(D+4)}\left(\widetilde{\mathcal{H}}_m(\boldsymbol{x})\right)^{-1/2}$$

where

$$\widetilde{\mathcal{H}}_m(\boldsymbol{x}) = \begin{cases} \mathcal{H}_m(\boldsymbol{x}) & \text{if} \quad \mathcal{H}_m(\boldsymbol{x}) \text{ is positive definite} \\ -\mathcal{H}_m(\boldsymbol{x}) & \text{if} \quad \mathcal{H}_m(\boldsymbol{x}) \text{ is negative definite} \end{cases}$$

(if $\mathcal{H}_m(\boldsymbol{x})$ is neither positive nor negative definite, the minimizer of AMSE($\widehat{m}(\boldsymbol{x};\boldsymbol{H})$) might not exist, even though the minimizer for the mean squared error exists). Hence, the matrix $\left(\widetilde{\mathcal{H}}_m(\boldsymbol{x})\right)^{-1/2}$ determines the shape and the orientation of the local optimal bandwidth region, while the term before it determines its overall magnitude, which depends among other factors on the sample size n, the dimension D of the covariate vector and on the integrated correlation function ρ_I. As in the independent error case, the optimal rate of convergence for the elements of $\boldsymbol{H}(\boldsymbol{x})$ is $O\left(n^{-1/(D+4)}\right)$.

The objective of most bandwidth selection methods is to obtain an estimator for the asymptotically optimal global bandwidth, defined as the minimizer of

$$\text{AMISE}\,(\boldsymbol{H}) = \int \text{AMSE}\,(\widehat{m}(\boldsymbol{x};\boldsymbol{H}))\, f(\boldsymbol{x})d\boldsymbol{x}.$$

Unlike in the local situation, no closed form expression is available for this bandwidth. In the next section, we will propose a bandwidth selection criterion based on Generalized Cross-Validation and that is asymptotically equivalent to AMISE$\,(\boldsymbol{H})$.

2.2 *Bandwidth Selection*

Consider selecting the bandwidth $\boldsymbol{H}$ that minimizes the GCV function

$$\text{GCV}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}_i}{1 - \frac{1}{n}\text{tr}(\boldsymbol{S})}\right)^2. \tag{6}$$

with $\boldsymbol{S}$ the $n\times n$ matrix whose ith row is equal to $\boldsymbol{s}^T_{\boldsymbol{X}_i}$, the smoother vector for $\boldsymbol{x} = \boldsymbol{X}_i$. Finding the minimizer of this function over the $D(D+1)/2$ parameters in $\boldsymbol{H}$ can be performed using numerical algorithms as implemented in statistical software, or with a specialized procedure such as the one proposed by Kauermann and Opsomer (2001). However, we cannot use this criterion directly for selecting the bandwidth in the presence of correlated errors, because its expectation is severely affected by the correlation. Specifically, Liu (2001) showed that the GCV is asymptotically biased, in the sense that

$$\text{E}\left\{\text{GCV}(\boldsymbol{H})|\,\mathcal{X}\right\} = \sigma^2 + \text{AMISE}(\boldsymbol{H}) - \frac{2\sigma^2 K(0)\rho_I}{n\,|\boldsymbol{H}|} + o_p\left(\lambda^4_{\max}(\boldsymbol{H}) + \frac{1}{n\,|\boldsymbol{H}|}\right).$$

Hence, when $\rho_I \neq 0$, the correlation contributes a term to the expectation of the GCV criterion that depends on $\boldsymbol{H}$ and is of the same order as the variance component of AMISE$(\boldsymbol{H})$ (see

(5)), which is likely to lead to inappropriate bandwidth choices (see Section 4 for an evaluation of this effect in practice).

We propose to remove this effect by using the "bias-corrected" GCV criterion

$$\mathrm{GCV}_c(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}_i}{1 - \frac{1}{n}\mathrm{tr}\,(\boldsymbol{S}\boldsymbol{R})}\right)^2, \tag{7}$$

with $\boldsymbol{R}$ the correlation matrix of the observations. Using the results of Liu (2001), it can be shown that $\mathrm{GCV}_c(\boldsymbol{H})$ is conditionally asymptotically unbiased, in the sense that

$$\mathrm{E}\left\{\mathrm{GCV}_c(\boldsymbol{H})|\,\mathcal{X}\right\} = \sigma^2 + \mathrm{AMISE}(\boldsymbol{H}) + o_p\left(\lambda^4_{\max}(\boldsymbol{H}) + \frac{1}{n\,|\boldsymbol{H}|}\right). \tag{8}$$

Also, under the additional assumption that $\mathrm{Var}(\frac{1}{n}\sum\varepsilon_i^2) = O(1/n)$, $\mathrm{GCV}_c(\boldsymbol{H})$ is consistent for $\mathrm{AMISE}(\boldsymbol{H})$ plus a term that does not depend on the bandwidth matrix $\boldsymbol{H}$. Expression (8) shows that, asymptotically, the bias in the criterion is indeed removed by the adjustment.

The criterion (7) is not yet a practical bandwidth selection criterion, since it requires knowledge of $\boldsymbol{R}$. Therefore, we will assume a parametric form for the correlation function, say $\rho_n(\cdot;\boldsymbol{\theta})$, and then replace the unknown $\boldsymbol{R}(\boldsymbol{\theta})$ in (7) by an estimate $\boldsymbol{R}(\hat{\boldsymbol{\theta}})$. We write the "bias-corrected and estimated" GCV criterion as

$$\mathrm{GCV}_{ce}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}_i}{1 - \frac{1}{n}\mathrm{tr}\left(\boldsymbol{S}\boldsymbol{R}(\widehat{\boldsymbol{\theta}})\right)}\right)^2. \tag{9}$$

How well (9) approximates (7) depends on whether the parametric form chosen for $\rho_n(\cdot;\boldsymbol{\theta})$ is appropriate, and how well $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$. The former aspect will be discussed in Section 4, while the latter is addressed in the following Result.

Result. 1. Assume that the parametric form of $\rho_n(\cdot;\boldsymbol{\theta})$ is correct, that $\rho_n(\cdot;\boldsymbol{\theta})$ is continuous and that its derivative(s) with respect to $\boldsymbol{\theta}$ are bounded. Assume also that $\boldsymbol{H}$ satisfies A3 and that there exists $\epsilon > 0$ such that $|1 - \frac{1}{n}\mathrm{tr}\,(\boldsymbol{S}\boldsymbol{R}(\boldsymbol{\theta}))\,| > \epsilon$ for all n. Under these assumptions,

$$\mathrm{GCV}_{ce}(\boldsymbol{H}) = \mathrm{GCV}_c(\boldsymbol{H}) + O_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|).$$

Under the stated assumptions, the result is immediate by an application of a first order Taylor series expansion for $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. The assumptions on $\rho_n(\cdot;\boldsymbol{\theta})$ are readily checked and hold for the exponential and rational quadratic models discussed above. While the assumption on $\frac{1}{n}\mathrm{tr}\,(\boldsymbol{SR}(\boldsymbol{\theta}))$ appears restrictive, note that it is also required for the GCV criterion itself to be a well-defined bandwidth selector. In particular, too small values for the diagonal elements of $\boldsymbol{H}$ could lead to violation of this assumption and would result in instability in the GCV (or GCV_{ce}) criterion. For a practical implementation of the GCV_{ce}, we would recommend that the assumption be checked with $\boldsymbol{\theta}$ replaced by its estimate and only those values for $\boldsymbol{H}$ satisfying the condition be allowed in the search for the minimizer. Putting a lower bound on the allowable values of $\boldsymbol{H}$ is common practice in GCV implementations.

The immediate consequence of Result 1 is that as long as $\boldsymbol{\theta}$ can be estimated consistently, then $\mathrm{GCV}_{ce}(\boldsymbol{H})$ is consistent for $\mathrm{GCV}_{c}(\boldsymbol{H})$. Therefore, it is also consistent for $\mathrm{AMISE}(\boldsymbol{H})$ and a term that does not depend on $\boldsymbol{H}$. In the next section, we discuss consistent estimators of the parameters of the correlation function for a specific function choice.

3. Estimation of the Correlation Function

The bias-corrected GCV criterion relies on the parametric specification of the correlation function. In this article, we consider the isotropic exponential model given in (3), but the approach will hold more generally for other models. It is also possible to extend the approach to other settings with anisotropy or non-zero nugget effects. The main requirement is that a parametric specification for the correlation function can be selected. We expect the exponential model to be broadly applicable for many spatial regression problems, since a positive correlation function that decreases smoothly in all directions very often provides a reasonable approximation for the observed spatial correlation. Because the focus of our approach is on estimating the mean function, not the correlation function, a high degree of accuracy in estimating the latter is not a prerequisite, and simple correlation models like the exponential will often suffice.

In order to estimate the spatial correlation function parameter(s), we express the correlation model as a semivariogram model, and estimate that model by an empirical semivariogram. For correlation function (3), the unknown parameter $\boldsymbol{\theta}$ that needs to be estimated under the semivariogram formulation is (σ^2, α). We will construct simple estimators for both quantities based on residuals $\widehat{\varepsilon}_i$ from a pilot fit using local linear estimator (2) and a pilot bandwidth matrix $\boldsymbol{\Lambda}_{pilot}$. The estimator for σ^2 is defined as

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2. \tag{10}$$

The estimator for α will be constructed from the empirical semivariogram

$$\widehat{\gamma}\,(d) = \frac{1}{2\,n(d,t)} \sum_{(i,j) \in S(d,t)} \left(\widehat{\varepsilon}_i - \widehat{\varepsilon}_j\right)^2, \tag{11}$$

(see Cressie, 1993, p.69), with $S(d,t) = \{(i,j) |\, d - t \leq \|\boldsymbol{X}_i - \boldsymbol{X}_j\| < d + t\}$ the set of all points that are at a distance $d \pm t$ of each other, and $n(d,t)$ the number of elements in $S(d,t)$. The number $t > 0$ is a tolerance value chosen to achieve a sufficient number of observations in $S(d,t)$. Under the assumptions made in this article, the empirical semivariogram $\widehat{\gamma}\,(\cdot)$ estimates the semivariogram function $\gamma_n(\cdot) = \sigma^2(1 - \rho_n(\cdot))$.

The $\widehat{\gamma}\,(d)$ can be calculated for any distance d. Here, we propose to calculate $\widehat{\gamma}\,(d)$ for a grid of distances $0 < d_1 < d_2 < \cdots < d_{J+1}$, and estimate α by

$$\widehat{\alpha}_J = \frac{1}{J} \sum_{j=1}^{J} \widehat{\alpha}_j, \tag{12}$$

where

$$\widehat{\alpha}_j = \frac{1}{nd_j} \left(\ln(\widehat{\sigma}^2) - \ln(\widehat{\sigma}^2 - \widehat{\gamma}(d_j))\right), \qquad j = 1, 2, \ldots, J. \tag{13}$$

Expression (13) is a "method of moment" estimator obtained by rewriting correlation model (3) as a semivariogram model and solving for α. Estimators at a set of J lags d_j are combined into (12) to improve the precision of the estimator. If a different correlation model than (3) is assumed, (13) will need to be replaced by a different corresponding expression. The following Results show that the estimators (10) and (11) are consistent. Proofs are in the Appendix.

Result. 2. Assume that the $\widehat{\varepsilon}_i$ are obtained from a pilot fit using local linear estimator (2) and a pilot bandwidth matrix $\boldsymbol{\Lambda}_{pilot}$. Let assumptions A1, A2 and A4 hold for the observations and the kernel, and let A3 hold after replacing $\boldsymbol{H}$ by $\boldsymbol{\Lambda}_{pilot}$. Then,

$$\widehat{\sigma}^2 = \sigma^2 + o_p(1).$$

Result. 3. Let the assumptions of Result 2 hold, and assume the first two derivatives of the semivariogram function and the density of the observed distances exist. Assume $\mathrm{E}(\varepsilon_i\varepsilon_j\varepsilon_k) = 0$ and $\mathrm{Cov}(\varepsilon_i\varepsilon_j, \varepsilon_k\varepsilon_l) = \mathrm{Cov}(\varepsilon_i, \varepsilon_k)\mathrm{Cov}(\varepsilon_j, \varepsilon_l) + \mathrm{Cov}(\varepsilon_i, \varepsilon_l)\mathrm{Cov}(\varepsilon_j, \varepsilon_k)$ for all i, j, k, l. Finally, assume that the tolerance value t satisfies $t \to 0$, $nt \to \infty$ as $n \to \infty$. Then, for any $d > 0$,

$$\widehat{\gamma}(d) = \gamma_n(d) + o_p(1).$$

Result 3 shows that under the stated assumptions, the empirical semivariogram is a consistent estimator for the semivariogram. The assumptions on the moments of ε_i are satisfied for the normal distribution and are made only to simplify the proof. They could be weakened and replaced by convergence bounds on moments of products of the errors, for instance. While the result is stated for $\hat{\gamma}(d)$ which is a function of residuals from a pilot fit, it clearly also holds if the errors themselves are observed. Hence, it shows the consistency of the empirical semivariogram to the true semivariogram for a stationary spatial random process. The method of proof in the Appendix treats the estimator $\widehat{\gamma}(d)$ as a kernel regression estimator with a uniform kernel $K(d) = I_{\{|d| \leq 1\}}$ and bandwidth parameter t. Using this approach resulted in a direct and simple proof of the consistency of the empirical semivariogram: after transforming the locations of the observations to polar coordinates, straightforward application of kernel asymptotics leads to the desired consistency. While we did not derive rates of convergence here, the same approach could certainly be used to do so.

Corollary. Under the assumptions of Result 3 and for any fixed J, we have

$$\hat{\alpha}_J = \alpha + o_p(1).$$

The corollary follows directly from Results 2 and 3 and the continuity of (13).

4. Simulation Experiments

In this section, a simulation study is carried out to evaluate the proposed bandwidth selection method. We compare criteria:

- GCV($\boldsymbol{H}$), given in (6),
- GCV$_c$($\boldsymbol{H}$), corrected for the true correlation function and given in (7),
- GCV$_{ce}$($\boldsymbol{H}$), corrected using an estimated correlation function assuming that the errors follow an exponential correlation model, and given in (9).

For this purpose, 200 samples of sample size $n = 400$ are generated following regression model (1), where the design points $\mathbf{X}_i = (X_{i1}, X_{i2})$ are uniformly sampled in the unit rectangle, the mean function is the additive model $m(X_1, X_2) = \sin(2\pi X_1) + 4(X_2 - 0.5)^2$ and the random errors ε_i are normally distributed with zero mean and exponential covariance function

$$Cov\left(\varepsilon_i, \varepsilon_j | \, \mathbf{X}_i, \mathbf{X}_j\right) = \sigma^2 \exp\left\{-a \left\|\mathbf{X}_i - \mathbf{X}_j\right\|\right\}, \tag{14}$$

with $\sigma = 0.4$. Several values of parameter a are considered: $a = 5$ (strong correlation), $a = 20$, $a = 40$ and $a = 200$ (approximately no correlation).

The local linear regression estimator (2) with Epanechnikov kernel $K(\mathbf{x}) = \frac{2}{\pi}\max\left\{\left(1 - \|\mathbf{x}\|^2\right), 0\right\}$ is used. For each simulated sample, we calculated the bandwidths obtained using the three

GCV criteria, denoted by $\mathbf{H}_{GCV}$, $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{GCVce}$ respectively. We will compare the methods based on their Mean Average Squared Error, MASE($\mathbf{H}$) $= \frac{1}{n}\left(\mathbf{Sm}-\mathbf{m}\right)^t\left(\mathbf{Sm}-\mathbf{m}\right) + \frac{1}{n}\sigma^2 \mathrm{tr}\left(\mathbf{SRS}^t\right)$, where $\mathbf{m} = \left(m\left(\mathbf{X}_1\right), m\left(\mathbf{X}_2\right), \ldots, m\left(\mathbf{X}_n\right)\right)^t$ and $\mathbf{R}$ is the true correlation matrix of the errors. For comparison, we also computed $\mathbf{H}_{opt}$, the minimizer of MASE($\mathbf{H}$).

In the computation of $\mathbf{H}_{GCVce}$, parameters σ^2 and a are estimated from nonparametric residuals obtained from a pilot fit with bandwidth $\mathbf{\Lambda}_{pilot} = \mathrm{diag}\{\hat{\sigma}_{X_1}, \hat{\sigma}_{X_2}\}$ where $\hat{\sigma}_{X_1}, \hat{\sigma}_{X_2}$ denote the sample standard errors of the two components of the $\boldsymbol{X}_i$. The distances $d_k = 0.001 + 2(k-1)t$, $k = 1, 2, \ldots, 30$ with $t = 0.005$ are used in the empirical semivariogram.

[Table 1 about here.]

Table 1 displays the MASE-optimal bandwidth $\mathbf{H}_{opt}$ and the simulation means of $\mathbf{H}_{GCV}$, $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{GCVce}$ as a function of a. When data are almost uncorrelated ($a = 200$), all three criteria produce similar bandwidths. The optimal bandwidth values increase as the correlation increases (a decreases). Bandwidths $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{GCVce}$ track this behavior, while $\mathbf{H}_{GCV}$ exhibit an oppositive behavior. For intermediate values of a, $\mathbf{H}_{GCVce}$ exhibits a slight tendency toward overestimation.

[Table 2 about here.]

Table 2 shows the simulated mean MASE corresponding to the three bandwidth selection methods and the minimum of MASE($\mathbf{H}$), i.e., MASE($\mathbf{H}_{opt}$). When we compare the methods based on their MASE values (Table 2), it is clear that $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{GCVce}$ are close to fully efficient, while the performance of $\mathbf{H}_{GCV}$ is unsatisfactory for values of $a \leq 40$.

We also considered the situation when the parametric covariance function is misspecified. The setup is the same as above, but the covariance model generating the errors follows rational quadratic covariance function:

$$Cov\left(\varepsilon_i, \varepsilon_j \middle| \mathbf{X}_i, \mathbf{X}_j\right) = \frac{\sigma^2}{1 + b\left\|\mathbf{X}_i - \mathbf{X}_j\right\|^2}, \tag{15}$$

where the dependence is controlled by parameter b. Two values of b are used: $b = 160$ and $b = 800$, corresponding to high and low levels of correlation. When estimating the covariances for the computation of $\mathbf{H}_{GCVce}$, mis-specified model (14) is used, while $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{opt}$ use correct model (15).

[Table 3 about here.]

[Table 4 about here.]

Table 3 displays the average values of $\mathbf{H}_{GCV}$, $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{GCVce}$, as well as the optimal bandwidth. Table 4 shows the mean MASE obtained using the bandwidths selected by three criteria, as well as MASE($\mathbf{H}_{opt}$). Although $\mathbf{H}_{GCVce}$ is not as successful as $\mathbf{H}_{GCVc}$ in adjusting for the correlation, both methods outperform $\mathbf{H}_{GCV}$. Hence, even in the situation when the covariance model cannot be specified exactly, the corrected and estimated GCV criterion proposed in Section 2.2 can provide an improvement over completely ignoring spatial correlation.

5. Example

We now return to the analysis of the Northeastern Lakes survey data. As mentioned in Section 1, the ANC measurements were collected on 338 lakes. Some lakes had repeated measurements, so that the total number of recorded ANC values is 557 and ranged from -72.2 to 3,371 μeq/L. Latitude and longitude of each lake centroid were recorded in decimal degree units, and the goal of the analysis is to produce a spatial surface for the mean ANC to visualize any broad spatial trends that are present over the region. Since we are not interested in the correlation structure of the data, we will not explicitly account for the fact that some lakes had multiple measurements and in particular, we will not include a nugget effect on the correlation structure. We implemented the local linear regression estimator as well as the GCV and corrected GCV procedures in `Matlab` 6.5.1.

[Figure 4 about here.]

Figure 4 shows the locations of the ANC measurements and the grid of locations on which local linear estimates of the mean will be calculated. The estimation grid was constructed by overlaying the survey region with a 50 × 50 grid and then dropping every grid point that did not satisfy one of the following two requirements: (a) it is within 1.5 "grid cell lengths" from an observation point, or (b) the calculation for the estimate at that grid point uses a smoothing vector (2) that is sufficiently stable. This second requirement was determined by performing a pilot local linear fit with an initial bandwidth matrix $\mathbf{\Lambda}_{pilot} = \text{diag}\{\hat{\sigma}_{X_{lat}}, \hat{\sigma}_{X_{long}}\} = \{1.445\ 0; 0\ 2.419\}$ and evaluating the matrix inversion in (2) using the `Matlab rcond()` function, with values above 0.05 considered acceptable. Both requirements are admittedly somewhat arbitrary, but they represent a compromise between coverage over the region of interest and ability to avoid singular design matrices. The resulting set of estimation points covers the interior of the ANC measurement locations, except for the very sparse region in the Western corner, where some points in the interior of the ANC measurement locations were dropped. The inability to deal with data sparseness is a drawback of kernel-based smoothing methods such as local linear regression. While local data sparseness could be handled by manually selecting larger bandwidths (locally or globally), we decided to instead avoid estimating at those "problem locations," since the goal of this article is to evaluate data-driven bandwidth selection methods.

Using the pilot bandwidth $\mathbf{\Lambda}_{pilot}$ as an initial value in the uncorrected GCV minimization procedure, the algorithm converged to

$$\boldsymbol{H}_{GCV} = \begin{bmatrix} 0.4067 & 0.001280 \\ 0.001280 & 1.024 \end{bmatrix}.$$

The resulting mean ANC surface evaluated on the estimation grid is displayed in Figure 2. As mentioned in Section 1, $\boldsymbol{H}_{GCV}$ is very close to the boundary of the feasible bandwidth region ("unfeasible" bandwidths are defined here as those that lead to a singular matrix for at least one of the local linear estimators (2) at the estimation points). This close proximity to the boundary results in highly unstable fits and convergence problems for the algorithm. The range of values obtained was from -26,310 to 32,065 μeq/L, well beyond the range of the ANC data, and numerous local minima and maxima occur in the estimated mean function.

For the correlation-corrected method proposed in this paper, we used the exponential correlation model (14) from Section 4. A pilot local linear regression using $\boldsymbol{\Lambda}_{pilot}$ was performed, resulting in parameter estimates $\hat{\sigma} = 496.1$ and $\hat{a} = 1.180$. When comparing the empirical semivariogram values with the semivariogram function that results from using $\hat{\sigma}$ and $\hat{a}$ in exponential model (14), we observed some non-random deviations between both sets of values. This could be due to the correlation of semivariogram estimates or to correlation model lack of fit. As discussed previously, however, precise estimation of the correlation function does not seem to be necessary for the purpose of correcting the GCV criterion, so that we proceeded with the exponential specification and these estimated values.

Starting from $\boldsymbol{\Lambda}_{pilot}$ again, the corrected GCV minimization converged to

$$\boldsymbol{H}_{GCVce} = \left[\begin{array}{cc} 1.200 & -0.000537 \\ -0.000537 & 1.793 \end{array} \right]$$

and the resulting estimated mean ANC surface is shown in Figure 3. In order to rule out a local minimum, we also performed the corrected GCV minimization using $\boldsymbol{H}_{GCV}$ as a starting value, but the algorithm converged to the same bandwidth value.

The fit in Figure 3 is much smoother than that in Figure 2, with values ranging from -369 to 2,387 μeq/L. The extreme low and high values occur at the boundary of the estimation region, where smoothing methods such as local linear regression often result in unreliable estimates. The lowest value is outside the range of observed ANC values, making it likely these very low estimates are indeed spurious. The high peak around coordinate values (44,-76) appears real, however, with all the observed ANC values above 2,000 indeed occurring in that area. In contrast, the estimated mean ANC surface using $\boldsymbol{H}_{GCV}$ displayed numerous local peaks above 2,000 that do not correspond to high observations.

In the analysis of ANC, the focus is most often on identifying areas where ANC values are low (200 μeq/L). The fit in Figure 3 clearly shows regions of lower ANC in the North and East portions of the map. Those same regions are visible in Figure 2, but that is mainly because we "processed" this map to use the same scale as Figure 3, which involved trimming all the peaks and valleys that fell outside of the range of the scale. In fact, having both fits agree with respect to the general location of the low ANC areas within the study region is a further indication that the observed pattern is actually present in the data, not just induced by bandwidth choice. Overall, correcting for correlation appears to have produced an estimated mean ANC surface that is not only smoother and computationally more stable, but also effectively visualizes potentially interesting patterns in the underlying data.

We performed a number of sensitivity analyses on the procedure and the obtained fit. First, we evaluated the effect of included a nugget effect in the correlation function. In the case of the exponential model, this implies that three parameters need to be estimated instead of two. The nugget effect is readily estimated from the repeated observations, and we adjusted the remaining parameter estimators accordingly. This had only a modest effect on the overall pilot fit. We obtained 3057.5 for the nugget effect, and the estimates for the remaining parameter were $\hat{\sigma} = 435.1$ and $\hat{a} = 1.528$. After adjusting the GCV criterion for the correlation function with nugget effect and plugging in the parameter estimates, we obtained a new bandwidth matrix $\{1.198 - 0.000159; -0.000159\ 1.796\}$, virtually indistinguishable from $\boldsymbol{H}_{GCVce}$ above.

Next, we wanted to investigate whether the correlation observed in the pilot fit residuals was in fact present in the data and not simply an "artifact" induced by our approach, since in the latter case, *any* dataset could potentially result in a substantial (and unwanted) GCV correction. To evaluate this, we performed a parametric bootstrap. We repeatedly resampled residuals from the fit obtained using $\boldsymbol{H}_{GCVce}$ and constructed bootstrap samples by randomly adding the resampled residuals to the model fits. Each bootstrap sample has approximately the same mean and variance function as the original dataset, but the errors are now spatially uncorrelated. We performed the pilot fit estimation of the correlation function parameters for each bootstrap sample, and the bootstrap mean value for the estimate of a was $\hat{a} = 8.399$, indicating a small amount of spatial correlation is indeed present in the residuals, even when the true errors are uncorrelated. However, when this $\hat{a}$ was used in the corrected GCV procedure, the bandwidth value obtained was $\{0.5625\ -0.000819; -0.000819\ 1.265\}$, close to $\mathbf{H}_{GCV}$ which assumes uncorrelated errors. Using these bandwidth in the local linear regression produced an estimated mean ANC surface that is visually very similar to that in Figure 2.

6. Conclusion

In this article, we proposed a simple GCV-based bandwidth selection method that is appropriate for smoothing spatial data when correlation is suspected. The method requires computation of a pilot fit and the specification of a parametric correlation function, so that the method is not totally "model-free." Nevertheless, the method appears reasonably robust to model misspecification, and works much better than completely ignoring the correlation and applying the traditional (unadjusted) GCV criterion. More generally, the approach described in this paper provides a framework to adjust data-driven smoothing parameter selection methods in the spatial context and can easily be extended by incorporating other smoothing methods, more complicated correlation function estimation methods, or other optimality criteria.

We demonstrated the applicability of the method on a dataset of ANC measurements for 338 lakes in the Northeastern U.S. In this data analysis, we showed that both the numerical stability of the algorithm and the resulting model fits improved dramatically after adjusting the bandwidth selection method for the presence of correlation, even when the spatial correlation function was imperfectly specified.

Acknowledgments

The research of M. Francisco-Fernández was partially supported by Grant PGIDIT03PXIC10505PN and MCyT Grant BFM2002-00265 (European FEDER support included). The research of J.D. Opsomer was partially supported by U.S. Environmental Protection Agency (EPA) STAR Research Assistance Agreement CR-829096. This manuscript has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report. The authors thank Phil Larsen, Alan Herlihy and an anonymous Associate Editor for helpful suggestions.

A. Proofs

Proof of Result 2: Write

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n \left(m\left(\boldsymbol{X}_i\right) - \widehat{m}(\boldsymbol{X}_i;\boldsymbol{\Lambda}_{pilot})\right)^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\left(\left(m\left(\boldsymbol{X}_i\right) - \widehat{m}(\boldsymbol{X}_i;\boldsymbol{\Lambda}_{pilot})\right)\right) \\ &= V_{n,1} + V_{n,2} + V_{n,3}.\end{aligned}$$

Using the ergodic theorem (Port, 1994, Proposition 59.9), $V_{n,1} \rightarrow \sigma^2$. Since the pilot bandwidth satisfies A3, (4) and (5) imply that

$$\mathrm{E}\left(\left(m\left(\boldsymbol{X}_i\right) - \widehat{m}(\boldsymbol{X}_i;\boldsymbol{\Lambda}_{pilot})\right)^2 |\mathcal{X}\right) = o_p(1), \tag{16}$$

and hence, using a straightforward extension of result 5 of Fuller (1996, p.300), $V_{n,2} = o_p(1)$. This immediately implies that also $V_{n,3} = o_p(1)$ by the Schwarz inequality.

■

Proof of Result 3: As in the proof of Result 2, write

$$\begin{aligned}\widehat{\gamma}\left(d\right) &= \frac{1}{2n(d,t)}\sum_{(i,j)\in S(d,t)}\left(\varepsilon_i - \varepsilon_j\right)^2 + \frac{1}{2n(d,t)}\sum_{(i,j)\in S(d,t)}\left(\left(\varepsilon_i - \widehat{\varepsilon}_i\right) - \left(\varepsilon_j - \widehat{\varepsilon}_j\right)\right)^2 \\ &\quad -\frac{1}{n(d,t)}\sum_{(i,j)\in S(d,t)}\left(\varepsilon_i - \varepsilon_j\right)\left(\left(\varepsilon_i - \widehat{\varepsilon}_i\right) - \left(\varepsilon_j - \widehat{\varepsilon}_j\right)\right) \\ &= \tilde{\gamma}(d) + \gamma_{n,2}(d) + \gamma_{n,3}(d),\end{aligned}$$

where $\tilde{\gamma}(d)$ is the empirical semivariogram of the errors. Since $\varepsilon_i - \widehat{\varepsilon}_i = \widehat{m}(\boldsymbol{X}_i; \boldsymbol{\Lambda}_{pilot}) - m(\boldsymbol{X}_i)$, we can use (16) and the approach of the proof of Result 2 to show that $\gamma_{n,2}(d), \gamma_{n,3}(d) = o_p(1)$. Hence, the result will be established if we prove that for any fixed $d > 0$,

$$\mathrm{E}\left((\widetilde{\gamma}(d) - \gamma_n(d))^2 | \mathcal{X}\right) = o_p(1). \tag{17}$$

We rewrite $\widetilde{\gamma}(d)$ as a Nadaraya-Watson estimator (see Wand and Jones, 1995, p.119)

$$\widetilde{\gamma}(d) = \frac{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{t} K\left(\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\| - d}{t}\right)(\varepsilon_i - \varepsilon_j)^2}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{t} K\left(\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\| - d}{t}\right)},$$

where the kernel function $K(d)$ is the indicator function $I_{\{|d| \leq 1\}}$.

To prove (17), we consider the bias and the variance components of the MSE separately. For the bias, we have

$$\begin{aligned}
\mathrm{E}(\widetilde{\gamma}(d) - \gamma_n(d) | \mathcal{X}) &= \gamma_n'(d) \frac{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{t} K\left(\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\| - d}{t}\right)(\|\boldsymbol{X}_i - \boldsymbol{X}_j\| - d)}{\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \frac{1}{t} K\left(\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\| - d}{t}\right)} (1 + o(1)) \\
&= \gamma_n'(d) \frac{A_1}{A_2} (1 + o(1)) = o_p(1)
\end{aligned}$$

by the differentiability assumptions for $\gamma_n(\cdot)$. The convergence in this result follows after we show that $\mathrm{E}(A_1^2) = o(1)$ and $\mathrm{E}(A_2 - df_d(d))^2 = o(1)$, where the expectation is now taken with respect to $\{\boldsymbol{X}_i, i = 1, \ldots, n\}$ and $f_d(\cdot)$ denotes the density of the distances. The summations in A_1^2 and $(A_2 - df_d(d))^2$ are over 4 indices, so that six cases need to be considered depending on which indices are taken to be equal. In all cases, a change to polar coordinates significantly simplifies the derivations:

$$\left\{ \begin{array}{rcl} \boldsymbol{X}_{i1} - \boldsymbol{X}_{j1} &=& r_{ij} \cos \theta_{ij} \\ \boldsymbol{X}_{i2} - \boldsymbol{X}_{j2} &=& r_{ij} \sin \theta_{ij}, \end{array} \right.$$

so that $\|\boldsymbol{X}_i - \boldsymbol{X}_j\| = r_{ij}$ The detailed calculations are omitted here.

For the variance component of the MSE in (17),

$$\begin{aligned}\operatorname{Var}\left(\widetilde{\gamma}\left(d\right)|\mathcal{X}\right) &= \frac{\frac{1}{n^4t^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n K\left(\frac{r_{ij}-d}{t}\right)K\left(\frac{r_{kl}-d}{t}\right)\operatorname{Cov}((\varepsilon_i-\varepsilon_j)^2,(\varepsilon_k-\varepsilon_l)^2)}{\frac{1}{n^4t^2}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n K\left(\frac{r_{ij}-d}{t}\right)K\left(\frac{r_{kl}-d}{t}\right)} \\ &= \frac{B_1}{B_2} = o_p(1).\end{aligned}$$

This is established by using the moment assumptions stated in Result 3, so that B_1 can be written as the sum of nine terms, say $B_{1g}, g = 1, \ldots, 9$, containing covariance products. For each term, it can be shown that $\mathrm{E}(|B_{1g}|) = o(1)$ (with respect to $f_x(\cdot)$), so that $B_1 = o_p(1)$ by a straightforward extension of result 5 of Fuller (1996, p.300). For B_2, we can show that $\mathrm{E}(B_2 - d^2 f_d(d)^2)^2 = o(1)$ by using the same approach as for A_2 above. Details are omitted.

■

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**, 749–759.

Chiu, S.-T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statistics and Probability Letters* **8**, 347–354.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, 2 edition.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*. John Wiley & Sons, New York, NY, 2 edition.

Hart, J. (1996). Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* **6**, 115–142.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B* **53**, 173–187.

Kauermann, G. and Opsomer, J. D. (2001). A fast method for implementing generalized cross-validation in multi-dimensional nonparametric regression. To appear in *Journal of Computational and Graphical Statistics*.

Larsen, D. P., Kincaid, T. M., Jacobs, S. E. and Urquhart, N. S. (2001). Designs for evaluating local and regional scale trends. *Bioscience* **51**, 1049–1058.

Liu, X. (2001). *Kernel smoothing for spatially correlated data*. PhD thesis, Department of Statistics, Iowa State University.

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

Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

Port, S. C. (1994). *Theorical Probability for Applications*. John Wiley & Sons, New York.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* **22**, 1346–1370.

Stoddard, J. L., Kahl, J. S., Deviney, F. A., DeWalle, D. R., Driscoll, C. T., Herlihy, A. T., Kellogg, J. H., Murdoch, P. S., Webb, J. R., and Webster, K. E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Technical Report EPA/620/R-03/001, U. S. Environmental Protection Agency, Washington, DC.

U.S. National Acid Precipitation Assessment Program (1991). 1990 Integrated Assessment Report. Technical report, Washington, DC.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wetzel, R. G. (1975). *Limnology*. W. B. Saunders Company, Philadelphia, PA.

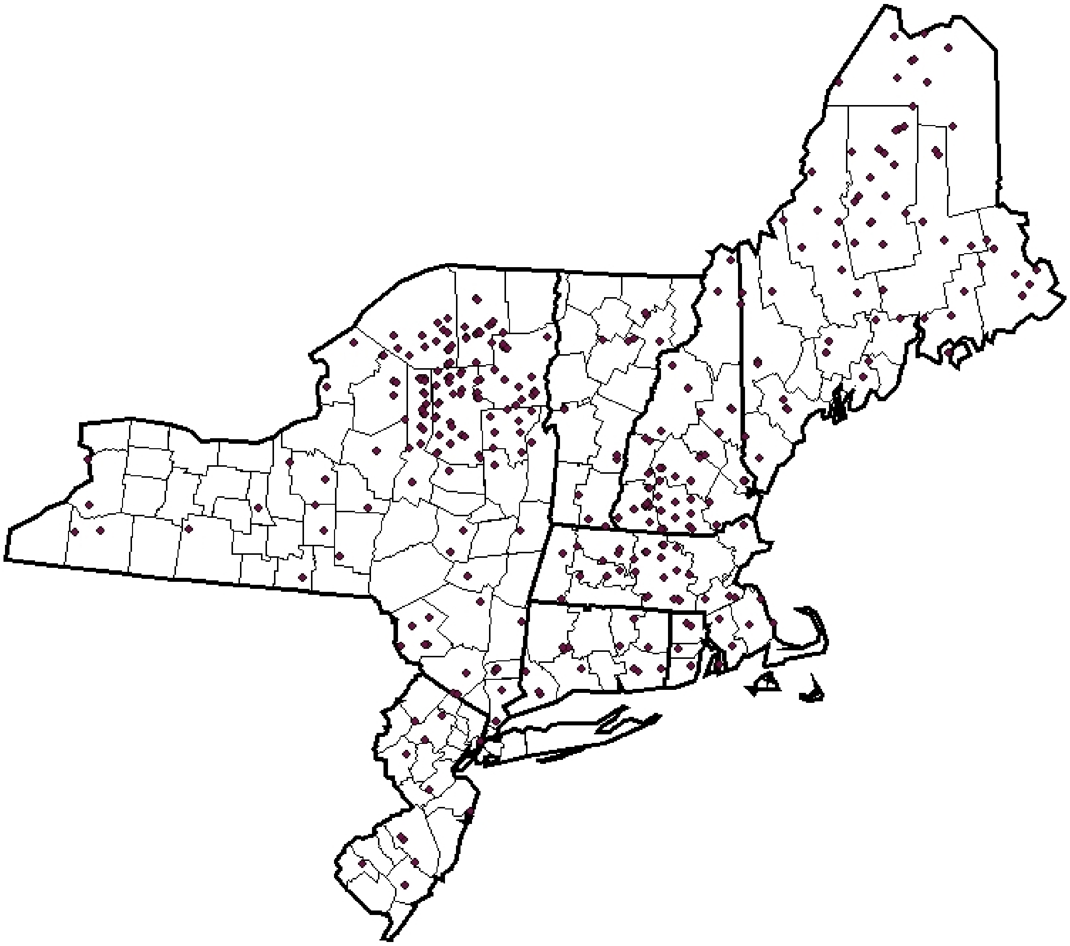

Figure 1. Map showing the location of sampled lakes in Northeastern U.S.

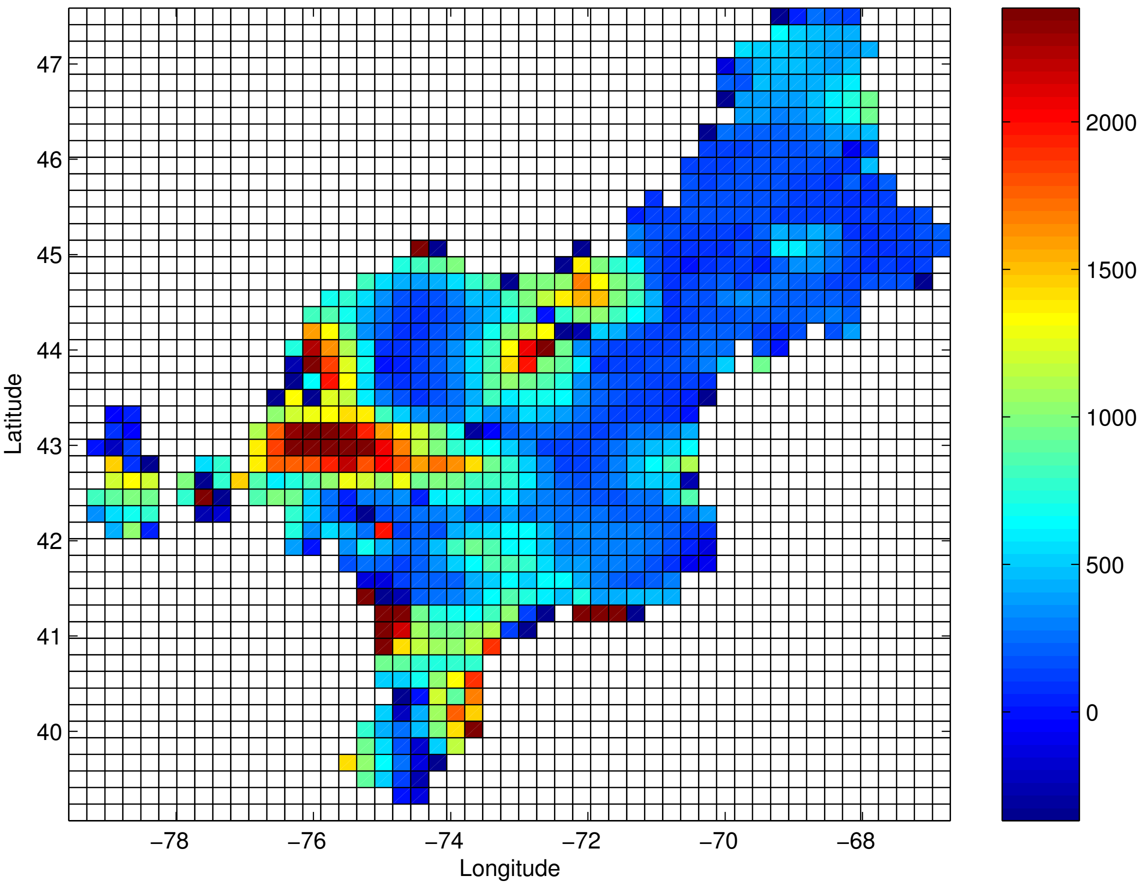

Figure 2. Local linear regression of ANC for Northeastern Lakes data, with bandwidth selected using GCV.

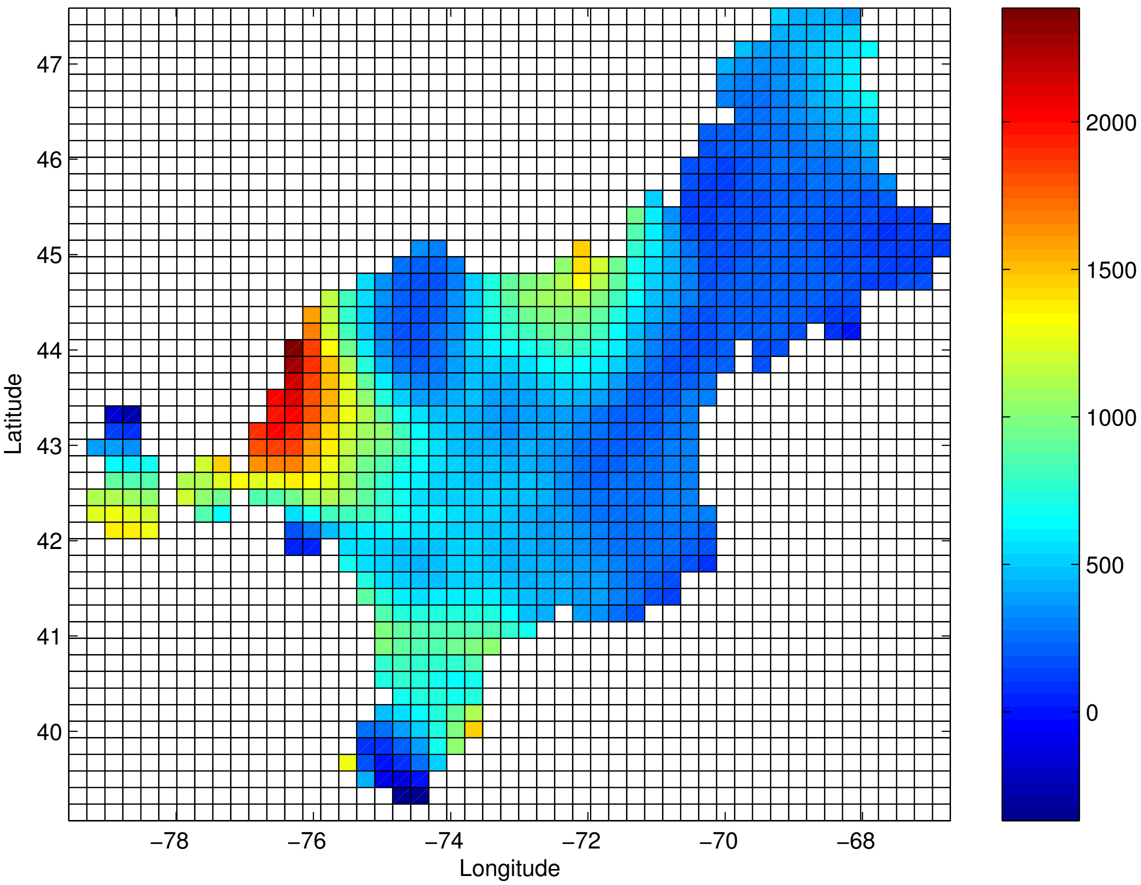

Figure 3. Local linear regression of ANC for Northeastern Lakes data, with bandwidth selected using GCV adjusted for spatial correlation.

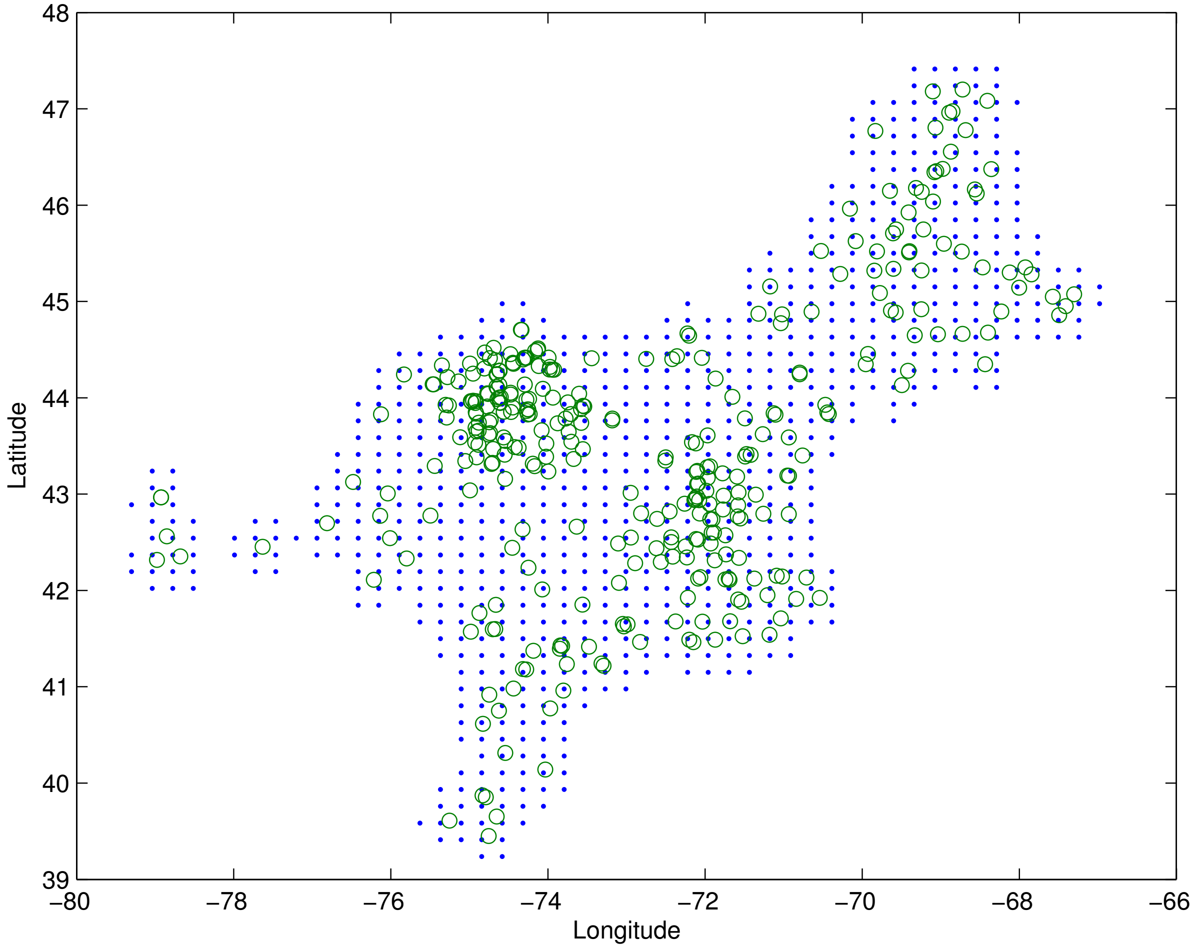

Figure 4. Locations (in latitude/longitude) of ANC measurements (o) and estimation grid (·) for Northeastern Lakes survey.

a	$\bar{\mathbf{H}}_{GCV}$	$\bar{\mathbf{H}}_{GCVc}$	$\bar{\mathbf{H}}_{GCVce}$	$\mathbf{H}_{opt}$
5	$\begin{pmatrix} 0.0837 & 0.0186 \\ 0.0186 & 0.0906 \end{pmatrix}$	$\begin{pmatrix} 0.1663 & 0 \\ 0 & 0.4077 \end{pmatrix}$	$\begin{pmatrix} 0.1638 & 0 \\ 0.0 & 0.4003 \end{pmatrix}$	$\begin{pmatrix} 0.1711 & 0 \\ 0 & 0.3731 \end{pmatrix}$
20	$\begin{pmatrix} 0.0988 & 0.0165 \\ 0.0165 & 0.0924 \end{pmatrix}$	$\begin{pmatrix} 0.1642 & 0 \\ 0 & 0.3350 \end{pmatrix}$	$\begin{pmatrix} 0.1806 & 0 \\ 0 & 0.4174 \end{pmatrix}$	$\begin{pmatrix} 0.1651 & 0 \\ 0 & 0.3246 \end{pmatrix}$
40	$\begin{pmatrix} 0.1013 & 0.0007 \\ 0.0007 & 0.1308 \end{pmatrix}$	$\begin{pmatrix} 0.1493 & 0.0102 \\ 0.0102 & 0.2969 \end{pmatrix}$	$\begin{pmatrix} 0.1683 & 0.0120 \\ 0.0120 & 0.3523 \end{pmatrix}$	$\begin{pmatrix} 0.1501 & -0.0013 \\ -0.0013 & 0.2964 \end{pmatrix}$
200	$\begin{pmatrix} 0.1390 & 0.0215 \\ 0.0215 & 0.2550 \end{pmatrix}$	$\begin{pmatrix} 0.1421 & 0.0194 \\ 0.0194 & 0.2650 \end{pmatrix}$	$\begin{pmatrix} 0.1526 & 0.0188 \\ 0.0188 & 0.2958 \end{pmatrix}$	$\begin{pmatrix} 0.1400 & -0.0033 \\ -0.0033 & 0.2641 \end{pmatrix}$

Table 1

Simulation means of $\mathbf{H}_{GCV}$, $\mathbf{H}_{GCVc}$ and $\mathbf{H}_{GCVce}$ and MASE-optimal bandwidth $\mathbf{H}_{opt}$, for correctly specified correlation function.

a	$\overline{\text{MASE}}(\mathbf{H}_{GCV})$	$\overline{\text{MASE}}(\mathbf{H}_{GCVc})$	$\overline{\text{MASE}}(\mathbf{H}_{GCVce})$	$\text{MASE}(\mathbf{H}_{opt})$
5	0.132032	0.096337	0.096095	0.092946
20	0.077291	0.037917	0.040116	0.036419
40	0.040352	0.021457	0.022587	0.020528
200	0.014468	0.014245	0.014600	0.013485

Table 2

Simulation means of $MASE(\mathbf{H})$ corresponding to three bandwidth selection method and $MASE(\mathbf{H}_{opt})$ for correctly specified correlation function.

b	$\bar{\mathbf{H}}_{GCV}$	$\bar{\mathbf{H}}_{GCVc}$	$\bar{\mathbf{H}}_{GCVce}$	$\mathbf{H}_{opt}$
160	$\begin{pmatrix} 0.0908 & 0.0198 \\ 0.0198 & 0.0960 \end{pmatrix}$	$\begin{pmatrix} 0.1743 & 0.0014 \\ 0.0014 & 0.3997 \end{pmatrix}$	$\begin{pmatrix} 0.2063 & 0.0013 \\ 0.0013 & 0,5944 \end{pmatrix}$	$\begin{pmatrix} 0.1783 & 0 \\ 0 & 0.3629 \end{pmatrix}$
800	$\begin{pmatrix} 0.0903 & 0.0149 \\ 0.0149 & 0.0882 \end{pmatrix}$	$\begin{pmatrix} 0.1592 & 0 \\ 0 & 0.3363 \end{pmatrix}$	$\begin{pmatrix} 0.2053 & 0 \\ 0 & 0.5639 \end{pmatrix}$	$\begin{pmatrix} 0.1620 & 0 \\ 0 & 0.3275 \end{pmatrix}$

Table 3

Simulation means of $\mathbf{H}_{GCV}$, $\mathbf{H}_{GCVc}$,$\mathbf{H}_{GCVce}$ and $\mathbf{H}_{opt}$ for mis-specified covariance function.

b	$\overline{\text{MASE}}(\mathbf{H}_{GCV})$	$\overline{\text{MASE}}(\mathbf{H}_{GCVc})$	$\overline{\text{MASE}}(\mathbf{H}_{GCVce})$	$\text{MASE}(\mathbf{H}_{opt})$
160	0.116272	0.069747	0.078452	0.065742
800	0.083326	0.035123	0.046823	0.033835

Table 4
Simulation means of MASE($\mathbf{H}$) corresponding to every bandwidth selection method and MASE($\mathbf{H}_{opt}$) for misspecified covariance function.