Variance Estimation in Spatial Regression Using a Non-Parametric Semivariogram Based on Residuals

Author(s): Hyon-Jung Kim and Dennis D. Boos

Source: *Scandinavian Journal of Statistics*, Sep., 2004, Vol. 31, No. 3 (Sep., 2004), pp. 387-401

Published by: Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics

Stable URL: https://www.jstor.org/stable/4616837

Variance Estimation in Spatial Regression Using a Non-parametric Semivariogram Based on Residuals

HYON-JUNG KIM
University of Oulu

DENNIS D. BOOS
North Carolina State University

ABSTRACT. The empirical semivariogram of residuals from a regression model with stationary errors may be used to estimate the covariance structure of the underlying process. For prediction (kriging) the bias of the semivariogram estimate induced by using residuals instead of errors has only a minor effect because the bias is small for small lags. However, for estimating the variance of estimated regression coefficients and of predictions, the bias due to using residuals can be quite substantial. Thus we propose a method for reducing this bias. The adjusted empirical semivariogram is then isotonized and made conditionally negative-definite and used to estimate the variance of estimated regression coefficients in a general estimating equations setup. Simulation results for least squares and robust regression show that the proposed method works well in linear models with stationary correlated errors.

Key words: estimating equations, Matérn family, restricted maximum likelihood, sandwich variance estimator

1. Introduction

Random processes over time or space typically have the property that nearby observations tend to be more alike than observations far apart. Scientific studies of such processes often involve modelling and estimation of a mean response with random errors assumed to be from a stationary process. For example, an additive error linear model is given by

$$Y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{x}_i$ may consist of variables that are functions of location as well as other covariates, and the e_i are errors whose correlation depends only on the distance between observations. Two applications of model (1) are as follows.

Example 1. For a project of the USDA Forest Service, the ozone index W126 (Lefohn & Runeckles, 1987) was desired at a number of sites in the north-eastern United States where plant damage on bio-indicator plants had been measured. W126 readings and some covariates were available for 174 other sites. Thus a regression model was developed between W126 and the following covariates: x_1 = elevation from sea level in metres, x_2 = mean annual rainfall in inches, x_3 = mean annual temperature in degrees Fahrenheit, x_4 = the Palmer drought severity index, and x_5 = the distance in metres to the closest city with a population over 50,000. We focus here on getting standard errors for the regression coefficients, but it was also of interest to get standard errors for the kriging predictions at the sites of the bio-indicator site monitors. The ordinary least squares (OLS) coefficients are *intercept*: −25.2 (23.5), x_1: .0063 (.003), x_2: −.26 (.29), x_3: 1.23 (.42), x_4: .60 (1.3), x_5: −.000064 (.00002). There is some evidence

of anisotropy. The standard errors from our proposed method are in parentheses and indicate that x_2 = rainfall and x_4 = Palmer drought severity may be dropped from the model.

Example 2. Global warming has been the focus of many studies, and is still a hotly debated topic. The data in Fig. 1, taken from Jones *et al.* (2000), are global mean annual temperatures for 1856–1999 with the 1961–90 mean subtracted. The data are overlaid with a simple OLS fit of the temperature deviations to x_1 = year − ymean and $x_2 = (\text{year} - \text{ymean})^2$, where ymean = 1927.5 is the mean of the years. The estimated linear and quadratic coefficients are .00425 and .0000421, respectively, with standard errors from our method of .000643 and .0000167. Semivariogram estimates of residuals from the first half and second half suggest a minor amount of covariance non-stationarity between the two halves; we ignore this possible non-stationarity in our analysis. As for any stationary error structure, these OLS coefficients are uncorrelated (due to centring the years), an approximate Wald test statistic for zero linear and quadratic coefficients is $(.00425/.000643)^2 + (.0000421/.0000167)^2 = 50.0$. Although a chi-squared distribution with 2 degrees of freedom may be a bit optimistic for the null distribution in this situation, the simulations in Table 3 suggest that 50.0 is highly significant.

One standard approach for model (1) with normally distributed errors is to assume a particular parametric semivariogram model, such as an exponential, and use restricted maximum likelihood (REML) for estimating the parameters of the semivariogram model and then to use estimated generalized least squares (EGLS) for estimating $\boldsymbol{\beta}$. Alternatively, one can avoid the normality assumption and use weighted least squares with the empirical semivariogram of the residuals to estimate the parameters of the semivariogram model (see Cressie, 1993, pp. 94–9, 165–70). In either case the variance of the EGLS $\widehat{\boldsymbol{\beta}}$ is estimated using standard formulas for generalized least squares with semivariogram estimates inserted where needed. A correct semivariogram model, however, is not always easy to choose, and we may be interested in estimation methods other than least squares.

The purpose of this paper is to give new methods for estimating the variance of $\widehat{\boldsymbol{\beta}}$ when $\widehat{\boldsymbol{\beta}}$ is obtained from a general estimating equations approach, i.e. satisfying $\sum \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}) = \mathbf{0}$, where $\boldsymbol{S}_i$ is usually a function of residuals and weights and $\widehat{\boldsymbol{\sigma}}$ are additional dispersion parameter estimates. This general class includes OLS (with $\boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}) = (Y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i$), EGLS, robust regression (e.g. Huber, 1980, Chapter 3, with $\boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}) =$

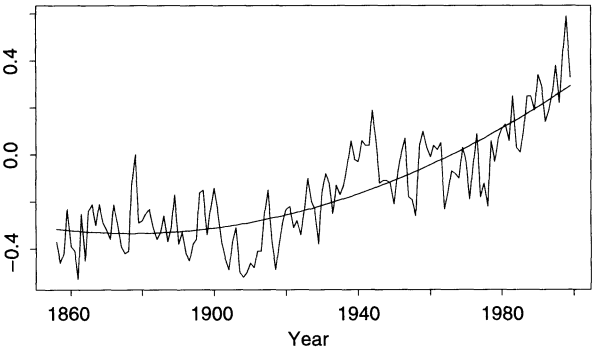

Fig. 1. Global mean annual temperatures for 1856–1999 with the 1961–90 mean subtracted; from Jones *et al.* (2000).

$(\psi[Y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}]/\widehat{\sigma})\boldsymbol{x}_i)$, and generalized estimating equations (GEE, Liang & Zeger, 1986). A parametric semivariogram model will not be chosen; rather, we will estimate the semivariogram subject only to a monotonicity constraint.

Our approach is similar in spirit to that of Lumley & Heagerty (1999), who use a weighted empirical variance estimate for the middle part of the 'sandwich' asymptotic variance formula in the estimating equations formulation. Their approach is more general because it allows non-stationary errors; in fact Lumley & Heagerty (1999) unify a variety of non-parametric methods including Newey & West (1987), White & Domowitz (1984), Andrews (1991), Lele (1991) and Yasui & Lele (1997). Although these estimators can be used for non-stationary models, one common problem is that they are not very efficient for highly correlated data (Andrews, 1991). Related non-parametric methods are found therein.

Our approach differs, though, in two ways from these other methods: we use non-parametric semivariogram estimation for the middle part of the 'sandwich', and we explicitly remove bias in the semivariogram estimates that would ordinarily result from using residuals rather than errors (which are of course unknown). Any non-parametric variance estimation method based on residuals that does not address the bias issue is doomed to underestimate the variance of estimated coefficients and of predictions. To our knowledge, no methods have successfully corrected the bias of residuals-based empirical variograms. Cressie (1993, pp. 167–8) notes that a kriging variance can be more influenced by the residuals-based variogram estimates. Here our main concern is with estimation of the variance of estimated regression parameters rather than kriging, and the substantial bias especially at moderate to large distances between observations (see Figure 3 in section 3.2) should not be ignored.

Our non-parametric semivariogram is based on several simple ideas. First, we construct a non-parametric semivariogram estimator in the constant mean case by monotonizing the standard empirical semivariogram (see section 3.1). We also check for positive-definiteness of the resulting covariance matrix and modify our estimate if it is not positive-definite. Next (section 3.2) we compute the bias of the residuals-based empirical semivariogram in model (1) when $\boldsymbol{\beta}$ is estimated by OLS. Then we correct for this bias by multiplying the empirical semivariogram by estimated factors computed at each distinct distance between observations. Finally, we monotonize and make the adjusted empirical semivariogram conditionally negative-definite. This non-parametric semivariogram is then used to estimate the variance of the estimated regression coefficients. Monte Carlo results for OLS estimates are given in section 4.1, and results for robust regression estimates are given in section 4.2. We begin in section 2 with a general explanation of how to estimate the variance of regression parameter estimates from data with correlated errors.

2. Variance estimation of regression parameter estimates

The estimating equations approach provides a general framework for deriving the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ that solves

$$\boldsymbol{G}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{0}$$

(and here for simplicity we have dropped the extra dispersion parameters σ mentioned in the Introduction). By Taylor series approximation,

$$\boldsymbol{0} = \boldsymbol{G}_n(\widehat{\boldsymbol{\beta}}) \approx \boldsymbol{G}_n(\boldsymbol{\beta}) + \dot{\boldsymbol{G}}_n(\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \boldsymbol{R}_n,$$

where

$$\dot{\boldsymbol{G}}_n(\widehat{\boldsymbol{\beta}}) = \frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \boldsymbol{G}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}).$$

Then, under suitable regularity conditions, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed with variance times n

$$A(\boldsymbol{\beta})^{-1} B(\boldsymbol{\beta}) \left\{ A(\boldsymbol{\beta})^{-1} \right\}^{\mathrm{T}} = \lim_{n \to \infty} [-\mathrm{E}\dot{\boldsymbol{G}}_n(\boldsymbol{\beta})]^{-1} \, n\mathrm{Var}[\boldsymbol{G}_n(\boldsymbol{\beta})] \left\{ [-\mathrm{E}\dot{\boldsymbol{G}}_n(\boldsymbol{\beta})]^{-1} \right\}^{\mathrm{T}}, \tag{2}$$

where it is assumed that there are matrices $A(\boldsymbol{\beta})$ and $B(\boldsymbol{\beta})$ such that $-\mathrm{E}\dot{\boldsymbol{G}}_n(\boldsymbol{\beta}) \to A(\boldsymbol{\beta})$ and $n\mathrm{Var}[\boldsymbol{G}_n(\boldsymbol{\beta})] \to B(\boldsymbol{\beta})$ as $n \to \infty$. In likelihood models $-\mathrm{E}\dot{\boldsymbol{G}}_n(\boldsymbol{\beta}) = n\mathrm{Var}[\boldsymbol{G}_n(\boldsymbol{\beta})]$ is the average Fisher information.

As the estimator

$$\dot{\boldsymbol{G}}_n(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})$$

will typically satisfy $\dot{\boldsymbol{G}}_n(\widehat{\boldsymbol{\beta}}) - \mathrm{E}\dot{\boldsymbol{G}}_n(\boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}$ even for correlated data, the problem of finding a consistent estimator of the asymptotic variance (2) is reduced to finding a consistent estimator for the middle term,

$$n\mathrm{Var}[\boldsymbol{G}_n(\boldsymbol{\beta})] = \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) \boldsymbol{S}_j(Y_j, \boldsymbol{x}_j, \boldsymbol{\beta})^{\mathrm{T}} \right].$$

Lumley & Heagerty (1999) point out that the empirical estimator

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) \boldsymbol{S}_j(Y_j, \boldsymbol{x}_j, \widehat{\boldsymbol{\beta}})^{\mathrm{T}} = \frac{1}{n} \left[\sum_{i=1}^{n} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) \right] \left[\sum_{j=1}^{n} \boldsymbol{S}_j(Y_j, \boldsymbol{x}_j, \widehat{\boldsymbol{\beta}}) \right]^{\mathrm{T}}$$

is identically zero by the definition of $\widehat{\boldsymbol{\beta}}$. This of course contrasts with the common situation where one can average over independent replications. Thus Lumley & Heagerty (1999) suggest estimating the middle term by

$$\widehat{J}_n(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ijn} \boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) \boldsymbol{S}_j(Y_j, \boldsymbol{x}_j, \widehat{\boldsymbol{\beta}})^{\mathrm{T}},$$

where $w_{ijn} \to 1$ as $n \to \infty$ but $w_{ijn} \to 0$ as the distance between two locations, $d(i, j) \to \infty$ for fixed n.

To implement our approach we first make the simplifying assumption that $\boldsymbol{S}_i$ has the form $\boldsymbol{S}_i(Y_i, \boldsymbol{x}_i, \boldsymbol{\beta}) = w_i(\boldsymbol{\beta}) S(e_i(\boldsymbol{\beta})) \boldsymbol{x}_i$, where S is now a real-valued function and the $e_i(\boldsymbol{\beta})$ are from a strictly stationary process. This form is found, for example, in OLS, EGLS, and robust regression. Thus, the $S(e_i(\boldsymbol{\beta}))$ are also from a strictly stationary process, and we will estimate the latter process using semivariogram techniques. We then employ it to estimate the middle term of (2) utilizing the relationship between semivariogram and covariance functions under second-order stationarity. We present the details of estimating the semivariogram in the next section.

3. Non-parametric semivariogram estimation

For an intrinsically stationary process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D, D \subset \mathbb{R}^d\}$, the semivariogram that characterizes its spatial dependence is defined by

$$\gamma(\boldsymbol{h}) = \frac{1}{2} \mathrm{Var}[Z(\boldsymbol{s} + \boldsymbol{h}) - Z(\boldsymbol{s})],$$

where $\boldsymbol{h}$ denotes the difference of locations. The function $2\gamma(\cdot)$ is called the variogram. For a sample of given realizations from $Z(\cdot)$, the empirical variogram is the unbiased estimator given by

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|}\sum_{N(\boldsymbol{h})}\{Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)\}^2,$$

where $N(\boldsymbol{h})$ is the set of distinct pairs whose spatial locations are separated by the distance $\boldsymbol{h}$. (We then use the simpler notation $\gamma(h)$ where h is the Euclidean distance between locations under isotropy.) Although the empirical variogram is unbiased for the variogram, it cannot be used directly in procedures such as kriging because it may not be conditionally negative-definite.

One standard approach to estimate the variogram is to choose a parametric variogram model (which by definition is conditionally negative-definite) and fit it by REML, maximum likelihood (ML), or weighted non-linear least squares (WNLS). There are a number of widely used parametric variogram models based on isotropic processes such as the exponential, spherical and Gaussian. The class (Matérn, 1960) appears to be the best choice of the present parametric models to estimate the dependence structure of a process as it includes or approximates a number of common models. The class can be defined by its isotropic autocovariance function:

$$C(h) = \frac{\sigma}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}h}{\rho}\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{2\nu^{1/2}h}{\rho}\right),$$

where σ is the scale parameter, ν the shape parameter, and ρ measures how quickly the correlation of the random field decays with distance. The function $\Gamma(\cdot)$ is the gamma function, and $\mathcal{K}_{\nu}$ is the modified Bessel function of the third kind of order ν. However, the underlying true variogram is rarely known in practice, and selection of a variogram model can be quite arbitrary.

There have been several attempts to avoid selecting variogram models via non-parametric variogram estimation (Shapiro & Botha, 1991; Hall *et al.*, 1994; Lele, 1995; Barry & Ver Hoef, 1996; Cherry *et al.*, 1996; Gorsich & Genton, 2000). Shapiro & Botha (1991) appear to be the first to have considered a non-parametric semivariogram estimate based on Bochner's theorem. Cherry *et al.* (1996) implemented the Shapiro–Botha estimator using the statistical package S-plus and found their non-parametric estimates performed well. One problem with their method is that the sill estimate tends to be biased and highly variable. Lele (1995) provided a non-parametric estimator of the semivariogram using a spline function and included a study of the performance of his estimator in terms of prediction and prediction error.

3.1. Monotone semivariogram estimation

Most parametric covariance models used in spatial analyses have correlations that are monotone decreasing with distance. Moreover, most physical processes exhibit this monotone behaviour as well. Thus, the centre-piece of our approach is to assume that the correlations are monotone decreasing with distance. This is a much weaker assumption than any of the common parametric semivariogram models.

Our basic approach to produce a non-parametric monotone semivariogram estimator is to apply the pool adjacent violators algorithm (PAVA, Barlow *et al.* 1972, p. 13) to empirical semivariogram estimates. Although the variance of the empirical semivariogram at a lag h is not exactly proportional to the inverse of the number of pairs, in the PAVA routine we follow Cressie (1993, p. 96) and use this weighting to reward the estimates at small lags.

The basic idea of PAVA in our context is the following: starting with $\gamma(h_1)$, move to the right and stop if $\gamma(h_i) > \gamma(h_{i+1})$. In that case replace $\gamma(h_i)$ and $\gamma(h_{i+1})$ by their weighted average $\gamma^*(h_i)$. Then move to the left to make sure that $\gamma(h_{i-1}) \leq \gamma^*(h_i)$ using the weighted average. Continue moving to the left and checking the monotone requirement, and then proceed again to the right. This process of averaging and back-averaging is continued until the right end point is reached.

As the variance of the empirical semivariogram is large for large lags, we follow the common practice of truncating the monotonized (also called isotonized) semivariogram at half the maximum distance found in the data set (Journel & Huijbregts, 1979). In order not to induce upward bias at that distance, we monotonize a larger set of lags before truncating. For example, for a time series on the equally spaced time points of 1,. . ., 100 we monotonize on the lags 1,. . ., 70 (number of pairs ≥30) and then truncate at 50.

Figure 2 gives averages of 1000 empirical and isotonized semivariograms for data on a 10×10 grid generated from an exponential model with range parameter = 1 (thus, the practical or effective range is 3) and sill = 3, $\gamma(h) = 3[1 - \exp(-h)]$. We can see that the empirical semivariogram is unbiased as advertised, and that the monotone estimate is somewhat biased downward in the middle. Table 1 shows, however, that the monotone estimate has lower mean squared error on the log scale for distances beyond 3. We use the log scale because the mean squared error for variance-like quantities is well-known to reward negatively biased estimators too much. Note that the standard errors of differences of entries in Table 1 are lower than the reported standard errors for individual entries.

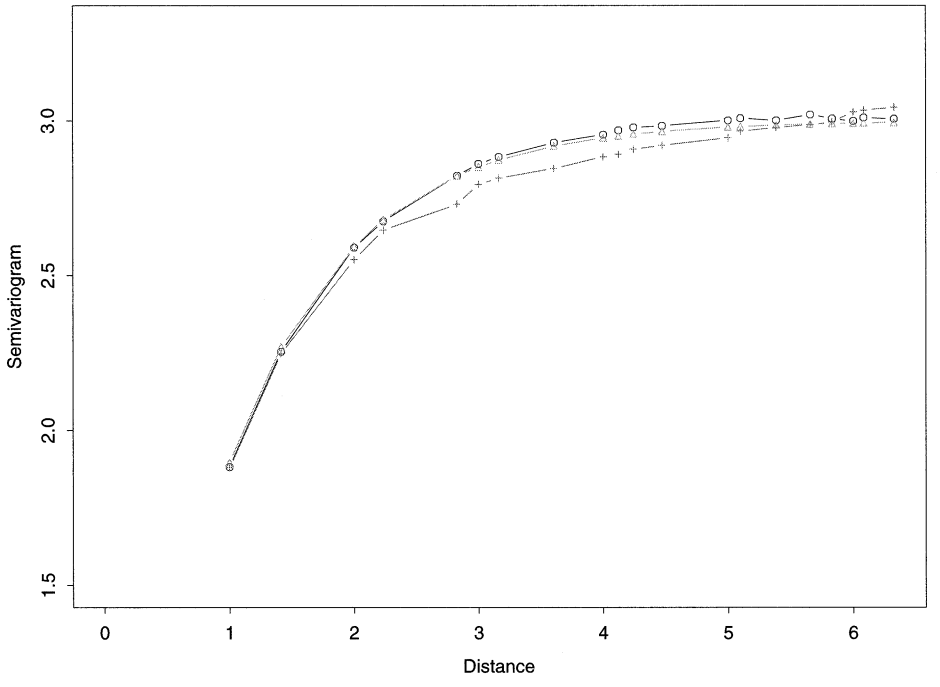

Fig. 2. Average of 1000 empirical semivariograms for data on a 10×10 grid generated from a constant mean process with an exponential (sill = 3, range = 3) semivariogram (△): empirical semivariogram (○) and isotonized version (+). Standard deviations of estimates are approximately bounded by .02.

Table 1. *Mean squared errors of the logarithm of empirical and monotone semivariogram estimates for constant mean normal data on a 10 × 10 grid with exponential semivariogram*

Distance	1.00	1.41	3.00	4.12	6.32
Range = 3, sill = 3					
Empirical	0.023	0.029	0.043	0.046	0.053
Monotone	0.023	0.028	0.039	0.040	0.043
Range = 6, sill = 3					
Empirical	0.022	0.030	0.060	0.078	0.110
Monotone	0.022	0.030	0.058	0.075	0.096
Average					
Standard errors	0.001	0.001	0.003	0.003	0.004

Results based on 1000 replications.

Very often, the isotonized semivariogram is already conditionally negative-definite. Sometimes it is not, and then we suggest using the spectral decomposition of the covariance matrix followed by replacement of negative eigenvalues by small positive ones (see Rousseeuw and Molenberghs, 1993, for other methods). For the variance estimation discussed in sections 4 and 5, this has virtually no effect. However, for other purposes such as kriging, it may be required. The isotonic estimators have a 'boxy' appearance, and sometimes it may be more appealing to smooth the isotonized semivariogram using a spline or other smoother. In our simulation studies (not presented here) we find that such additional smoothing makes very little difference in variance estimates on average. Consistency of the monotonized empirical semivariogram follows simply from the consistency of the empirical semivariogram as long as the underlying semivariogram is monotone.

Theorem 1.
Suppose that the empirical variogram $\widehat{\gamma}(\mathrm{h})$ *is strongly consistent for a non-decreasing semivariogram* $\gamma(h)$ *at the points* $h_1, \ldots, h_k$. *Then the monotonized empirical variogram is also strongly consistent for* $\gamma(h)$ *at the same points.*

Proof. For simplicity, consider the case of a strictly increasing semivariogram. Then for a sufficiently large sample size with probability 1, $\widehat{\gamma}(h_1) < \widehat{\gamma}(h_2) < \cdots < \widehat{\gamma}(h_k)$, and the monotonized version is the same as the empirical semivariogram. This same idea extends to cases that are not strictly monotone.

3.2. Bias correction of residuals-based semivariogram

Consider model (1) where the errors e_i are drawn from a mean zero, second-order stationary random error process. We shall assume that the unknown regression coefficients $\boldsymbol{\beta}$ are estimated by OLS. A natural approach is then to construct the empirical semivariogram from the residuals. Unfortunately, the distribution of the residuals is not the same as the errors, and the empirical semivariogram based on residuals is seriously biased downward. Figure 3 shows the average of the empirical semivariogram for 1000 replications of a constant mean normal process with an exponential semivariogram having range = 3 and sill = 3 for a 10 × 10 grid. The middle curve is for residuals from (1) with errors from this process based on fitting an intercept and the location coordinates (x and y). The lower curve is based on the residuals from a fit when the mean function is the same three variables and three more independent randomly generated standard normal variables.

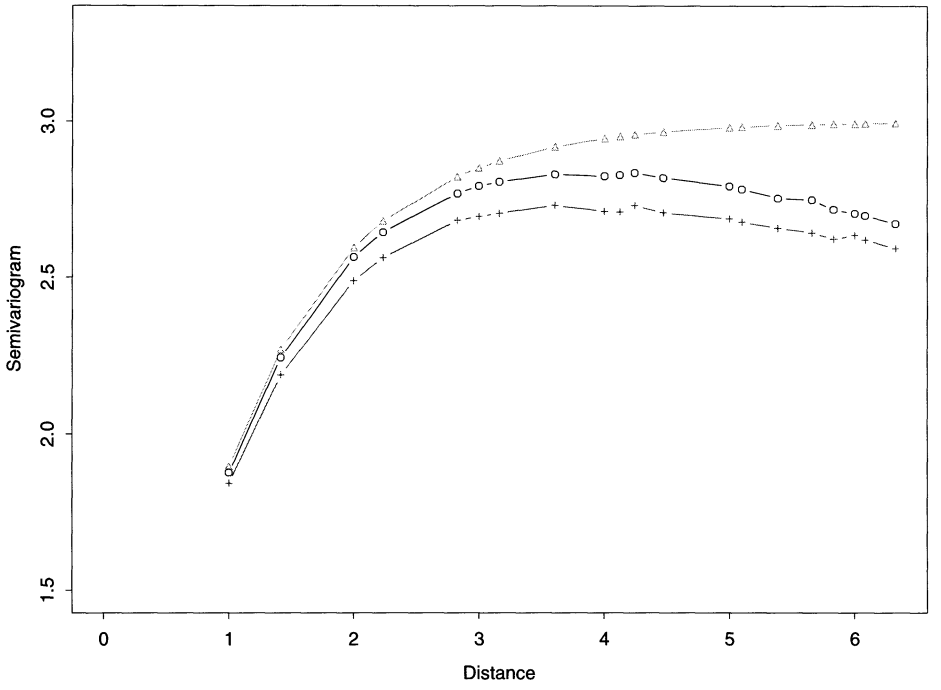

Fig. 3. Exponential (sill = 3, range = 3) semivariogram (△) for data on a 10 × 10 grid, average of 1000 empirical semivariograms from residuals with $p = 3$ estimated parameters (○) and with $p = 6$ estimated parameters (+). Standard deviations of estimates are approximately bounded by .02.

To understand more clearly the effect of the residuals, let V denote the covariance matrix of the errors for a sample with n data points. Then simple calculations show that the covariance matrix of the OLS residuals $\hat{e}_i = Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is given by $(I - P)V(I - P)$, where I is the n-dimensional identity matrix and $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the projection matrix of X, where X is formed from the row vectors $\boldsymbol{x}_i^{\mathrm{T}}$.

The expected value of the empirical semivariogram of the residuals at lag h is

$$\begin{aligned}\mathrm{E}_{\mathrm{res}}(h) &= \frac{1}{2|N(h)|}\sum_{N(h)}\mathrm{E}(\hat{e}_{i+h} - \hat{e}_i)^2 \\ &= \frac{1}{|N(h)|}\sum_{N(h)}(I - P)V(I - P)_{i,i} - \frac{1}{|N(h)|}\sum_{N(h)}[(I - P)V(I - P)]_{i,i+h}.\end{aligned}$$

The expected value of the empirical semivariogram of the errors at lag h is of course $\gamma(h)$, but we can write it in a form similar to the above:

$$\mathrm{E}_{\mathrm{err}}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}\mathrm{E}(e_{i+h} - e_i)^2 = \sigma^2 - \frac{1}{|N(h)|}\sum_{N(h)}[V]_{i,i+h},$$

where σ^2 is the variance of the errors. The ratio of these quantities, $\mathrm{fac}(h) = \mathrm{E}_{\mathrm{err}}(h)/\mathrm{E}_{\mathrm{res}}(h)$, is essentially the ratio of the true semivariogram $\gamma(h)$ to the average of the empirical semivariograms in Fig. 3.

We feel that the downward bias in Fig. 3 is unacceptable. Moreover, applying PAVA to the residuals-based empirical semivariogram will certainly improve it, but the scope for

improvement is quite limited. Thus we feel it is important first to adjust the empirical semivariogram before monotonizing. Our approach then is to estimate fac(h) by using estimated covariances $\widehat{V}$ obtained from the monotonized version of the residuals-based empirical semivariogram. We multiply the original residuals-based empirical semivariogram by these estimated factors resulting in a bias-reduced empirical semivariogram. Finally we monotonize this bias-reduced empirical semivariogram to obtain our estimated semivariogram. It is also possible to iterate the factor estimation step, but we found that it did not make much difference. Figure 4 shows the average of the residuals-based empirical semivariogram, the bias-reduced empirical semivariogram, and the monotonized version for 1000 samples from a linear model with intercept and locations fitted ($p = 3$ case) and errors generated from an exponential semivariogram with range = 6 and sill = 3. Comparing Fig. 4 with Fig. 2, one can see that monotonizing has a similar effect; that is, it tends to pull down on average the bias-reduced empirical semivariogram.

Table 2 shows that the monotonized version has improved mean squared error properties relative to the original residuals-based empirical semivariogram.

Figure 5 shows that a parametric semivariogram estimated by WNLS fitted to the residuals-based empirical semivariogram has considerable bias. The average values of the range parameter and sill using the Monte Carlo results that produced Fig. 5 are .85 and 2.82, respectively (true values are 1.0 and 3.0). If instead one uses the bias-reduced empirical semivariogram, then the WNLS method produces nearly unbiased results (not shown in Fig. 5): the corresponding average of range parameter and sill estimates are 1.01 and 3.03, respectively.

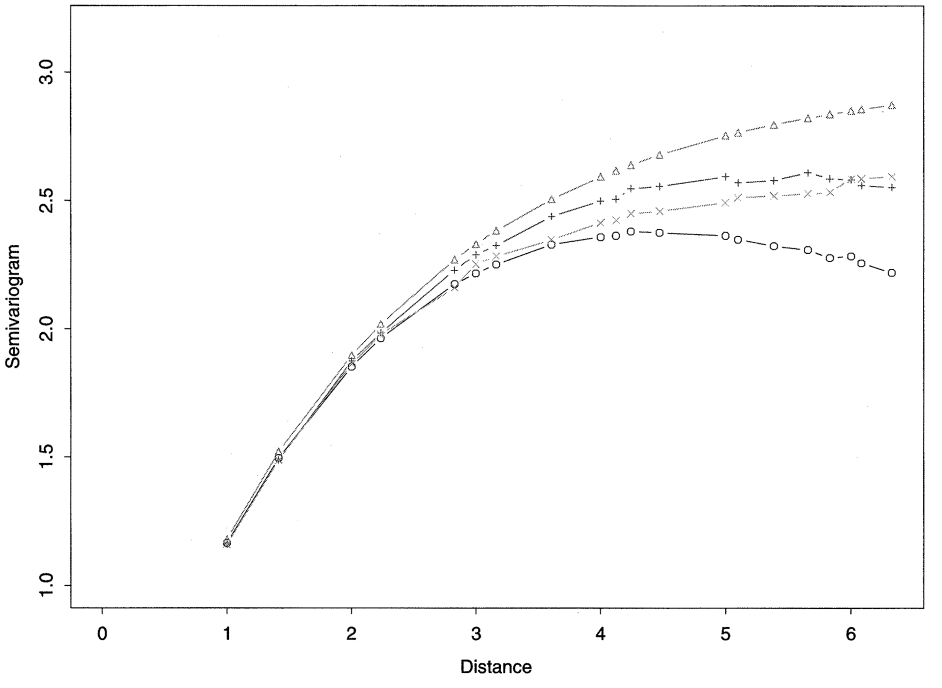

Fig. 4. Exponential (sill = 3, range = 6) semivariogram (△) on a 10 × 10 grid and the average of 1000 replications of semivariogram estimates based on residuals from $p = 3$ estimated parameters: empirical semivariogram from residuals (○), bias-reduced empirical semivariogram (+), monotonized bias-reduced empirical semivariogram (×). Standard deviations of estimates are approximately bounded by .02.

Table 2. *Mean squared errors of logarithms of semivariogram estimates: empirical, bias-reduced empirical and monotone bias-reduced*

Distance	1.00	1.41	3.00	4.12	6.32
Range = 3, sill = 3					
Empirical	0.023	0.029	0.046	0.052	0.067
Bias-reduced	0.023	0.029	0.046	0.051	0.060
Monotone bias-reduced	0.023	0.029	0.042	0.045	0.047
Range = 6, sill = 3					
Empirical	0.023	0.032	0.069	0.098	0.173
Bias-reduced	0.023	0.032	0.067	0.092	0.139
Monotone bias-reduced	0.023	0.032	0.067	0.091	0.116
Average					
Standard errors	0.001	0.001	0.003	0.003	0.004

Data on a 10 × 10 grid from exponential semivariogram.
Results based on 1000 replications.

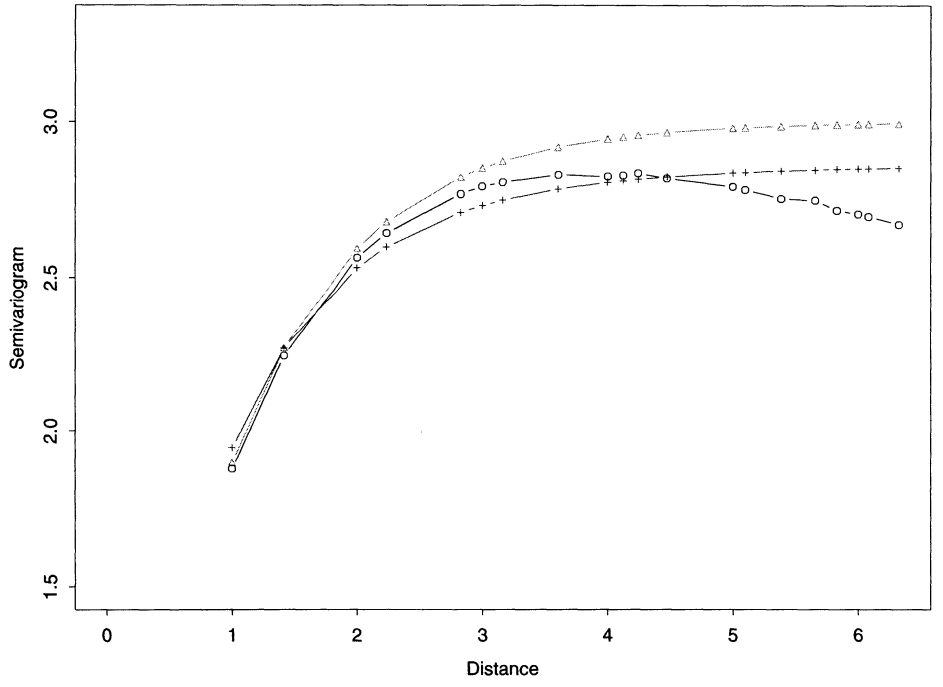

Fig. 5. Exponential (sill = 3, range = 3) semivariogram (△), average of 1000 replications of semivariogram estimates based on residuals from $p = 3$ estimated parameters: empirical semivariogram from residuals (○) and weighted nonlinear least squares parametric estimates (+). Standard deviations of estimates are approximately bounded by .02.

In general, non-parametric semivariogram methods will inherit the bias problem of the empirical semivariogram at large lags. We have not seen attempts to correct this bias directly as we suggest here, but Matheron (1973) and Cressie (1987) propose a method similar to differencing in time series for handling the bias issue. For visual confirmation of parametric

Table 3. *Average of standard deviation estimates (divided by true standard deviation) for regression from time series data of size 100 with autoregressive errors*

	ρ	Intercept	Trend	Season
Monotone ESV	0.1	1.10 (0.036)	1.08 (0.031)	1.04 (0.019)
WEAVE		0.97 (0.009)	0.95 (0.016)	0.94 (0.013)
Monotone ESV	0.5	1.19 (0.051)	1.15 (0.042)	1.00 (0.026)
WEAVE		0.87 (0.015)	0.85 (0.018)	0.84 (0.018)
Monotone ESV	0.9	0.97 (0.054)	0.96 (0.048)	0.84 (0.033)
WEAVE		0.54 (0.021)	0.53 (0.022)	0.65 (0.026)

Results based on 100 replications. Standard errors are in parentheses.

Table 4. *Average of standard deviation estimates (divided by true standard deviation) for regression from 10 × 10 spatial data of size 100 with exponential variogram errors*

	Range	Sill	Intercept	x coordinate	y coordinate
Monotone ESV	3	3	0.89 (0.033)	0.90 (0.034)	0.90 (0.034)
REML (Matérn)			0.86 (0.031)	0.86 (0.025)	0.86 (0.025)
REML (spherical)			0.82 (0.030)	0.83 (0.014)	0.83 (0.014)
REML (Gaussian)			0.65 (0.010)	0.66 (0.005)	0.66 (0.005)
WNLS (exponential)			0.87 (0.025)	0.87 (0.024)	0.87 (0.024)
Monotone ESV	6	3	0.73 (0.029)	0.76 (0.030)	0.76 (0.030)
REML (Matérn)			0.71 (0.025)	0.75 (0.024)	0.75 (0.024)
REML (spherical)			0.69 (0.045)	0.72 (0.015)	0.72 (0.015)
REML (Gaussian)			0.42 (0.010)	0.44 (0.005)	0.44 (0.005)
WNLS (exponential)			0.73 (0.029)	0.74 (0.026)	0.74 (0.026)

Results based on 100 replications. Standard errors are in parentheses.

Table 5. *Average of standard deviation estimates (divided by true standard deviation) for regression from 10 × 10 spatial data of size 100 with t-transformed exponential variogram errors.*

	Range	Sill	Intercept	x coordinate	y coordinate
True standard deviation (LS)	3	3	0.84	0.118	0.118
True standard deviation (robust)			0.75	0.106	0.106
Monotone ESV (LS)	3	3	0.91 (0.036)	0.91 (0.034)	0.91 (0.034)
Monotone ESV (robust)			0.89 (0.035)	0.90 (0.032)	0.90 (0.040)
True standard deviation (LS)	6	3	1.25	0.168	0.168
True standard deviation (robust)			1.14	0.154	0.154
Monotone ESV (LS)	6	3	0.73 (0.031)	0.75 (0.031)	0.75 (0.031)
Monotone ESV (robust)			0.73 (0.032)	0.76 (0.032)	0.77 (0.036)

Results based on 100 replications. Standard errors are in parentheses. True standard deviations are based on 10,000 replications and have estimated coefficient of variation .007.

variogram models, Brownie & Gumpertz (1997) suggest adjusting REML estimates of parametric semivariograms so that graphically they are aligned with a plot of the empirical semivariogram of residuals.

The estimation approach least likely to be affected by residual bias is REML. REML maximizes the likelihood of error contrasts that does not depend on estimated regression parameters, and does not suffer from the often severe underestimation of the parameters found with regular ML. However, one disadvantage of likelihood estimation procedures is

that they rely on the Gaussian assumptions, and such assumptions are often inappropriate; the underlying distribution of most processes is not known or contamination of the distribution can occur by a few errant observations. Non-parametric estimation of the variogram may provide less-sensitive estimates for some non-Gaussian models. Of course, least squares estimation of regression coefficients is also questionable in the face of non-Gaussian errors. In section 4.2, we consider using robust regression estimates.

4. Variance estimation

4.1. Ordinary least squares estimates of $\boldsymbol{\beta}$

Under the model (1) with second order stationary errors and $\boldsymbol{\beta}$ estimated by OLS, the variance of $\widehat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}VX(X^{\mathrm{T}}X)^{-1},$$

where V is the covariance matrix of the errors. For estimating this variance, we utilize the relationship between a semivariogram function $\gamma(h)$ and a covariance function $C(h)$ under second-order stationarity, $\gamma(h) = C(0) - C(h)$ and $C(h) = \gamma(\infty) - \gamma(h)$, $C(0) = \gamma(\infty)$ is the variance of the errors. Then we use the monotone semivariogram estimates proposed in section 3.2 to estimate V and substitute in the above expression.

Recall that Fig. 4 shows how the correlations tend to be overestimated by our monotone semivariogram estimate. This happens because on average in the middle distances the monotone estimate is biased downward. On a sample by sample case, this can be seen when the middle part of the monotone estimate is flat but rises near the half maximum distance. In such a case, the correlations are positive for a large number of distances h. This results in variance estimates that can be too large. Thus we use a cut-off rule for correlations: we set any estimated correlation to 0 if its value is smaller than $1/\sqrt{n}$. This is a fairly arbitrary rule, but it is similar in spirit to the WEAVE weight functions of Lumley & Heagerty (1999). Basically, the number of non-zero correlations cannot be growing too quickly with n if we are to have good variance estimates.

Table 3 shows results for errors from an ar(1) process with $\rho = .1, .5$ and $.9$. The X matrix consists of an intercept term, a trend (from -1 to 1 increasing linearly), and a seasonal term (a cosine wave with two cycles), which were taken from the simulation study of Lumley & Heagerty (1999, p. 469). For comparison, we also report results using the approach of Lumley & Heagerty (1999) denoted by 'WEAVE' in Table 3. We obtained their results using a program from their website with default values. Perhaps other settings would have performed better for the $\rho = .9$ case. Our results, denoted by 'Monotone ESV' for monotone empirical semivariogram, are reasonably on target (values close to 1.00), but more variable than the WEAVE results. Part of that higher variability can be explained by the fact that our estimates are larger on average than the WEAVE estimates.

Table 4 shows results for Gaussian spatial processes observed on a 10×10 grid, that are generated using an exponential variogram model with sill = 3 and practical range = 3 and 6, respectively. The average correlation is 0.04 for the model with the practical range equal to 3, and is 0.14 for the model with the practical range 6. (For comparison with the time series simulation in Table 3, note that the average correlations for $\rho = .1, .5$ and $.9$ are .002, .02 and .16, respectively.) A mean surface was added to these Gaussian processes to describe a spatial trend: $f(\mu) = \beta_0 + \beta_1 x + \beta_2 y$, where (x, y) defines a point in $\mathbb{R}^2$ with $\beta_0 = .0$, $\beta_1 = .9$, $\beta_2 = .06$. Three covariates, a column of ones for the intercept and x (longitude) and y (latitude), were fitted to the model to estimate the mean surface. We then compared our methods with the REML estimation using an assumed Matérn

variogram (without nugget, which is correct here) and also with several incorrect models, spherical and Gaussian. We see in Table 4 that our estimates are a little closer to the target value and a little more variable than other estimates. The estimates under the competing misspecified models such as spherical and Gaussian models yield relatively poor results as expected.

4.2. Robust regression estimates of $\boldsymbol{\beta}$

The OLS estimators minimize the sum of residual squares $\sum_{i=1}^{n} e_i^2(\boldsymbol{\beta})$, and efficiency losses may arise as this sum of squares is sensitive to large values that occur more frequently with non-Gaussian data. Robust M-estimators minimize an objective function $\sum_{i=1}^{n} \rho(e_i(\boldsymbol{\beta}))$ that is less sensitive to large values. Equivalently by taking derivatives, one solves $\sum_{i=1}^{n} \psi(e_i(\boldsymbol{\beta}))\boldsymbol{x}_i = 0$, where $\psi = \rho'$. In order for the estimators to be scale invariant, one needs to alter the above equation to $\sum_{i=1}^{n} \psi(e_i(\boldsymbol{\beta})/\widehat{\sigma})\boldsymbol{x}_i = 0$, where $\widehat{\sigma}$ is a scale estimate. There are different options for selecting ρ and $\widehat{\sigma}$, but a common approach is to use

$$\rho(t) = \begin{cases} t^2/2 & \text{if } |t| \le k, \\ k|t| - k^2/2 & \text{otherwise,} \end{cases} \quad \psi(t) = \begin{cases} -k & \text{if } t < k, \\ t & \text{if } -k \le t \le k, \\ k & \text{if } t > k, \end{cases}$$

where k is a tuning constant, usually selected to give an appropriate asymptotic efficiency when the data are Gaussian. 'Huber's proposal 2' finds the scale estimate $\widehat{\sigma}$ by solving simultaneously $\sum_{i=1}^{n} \psi^2(e_i(\boldsymbol{\beta})/\widehat{\sigma}) = C_k$, where C_k is a constant chosen so that $\widehat{\sigma}$ is consistent for the standard deviation when the data are from a Gaussian distribution. For example, when $k = 1$, $C_1 = .516$, and the asymptotic relative efficiency of $\widehat{\boldsymbol{\beta}}$ to least squares for independent Gaussian data is close to 90%.

Asymptotic normality of $\widehat{\boldsymbol{\beta}}$ was proved by Koul (1977, p. 688). One version of the asymptotic variance is given by

$$D(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}V_{\psi}X(X^{\mathrm{T}}X)^{-1},$$

where $D = \sigma/\mathrm{E}\psi'(e_1/\sigma)$ and V_{ψ} is the covariance matrix of $\psi(e_1/\sigma),\ldots, \psi(e_n/\sigma)$. We should also mention that we have ignored the role of estimating σ in the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ because for errors with symmetric marginal distributions, the asymptotic correlation between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}$ is zero.

These M-estimators fit into the general scheme outlined in section 2. That is, we assume that $\psi(e_1/\sigma),\ldots, \psi(e_n/\sigma)$ are second-order stationary and estimate their covariance matrix using our non-parametric semivariogram estimator based on treating $\psi(\widehat{e}_1/\widehat{\sigma}),\ldots,\psi(\widehat{e}_n/\widehat{\sigma})$ as residuals from least squares regression. The estimated factors for bias reduction are not quite correct as they officially only apply to least squares residuals. However, one way to motivate their use is by defining pseudo-observations $\widetilde{Y}_i = \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + d\psi(\widehat{e}_i/\widehat{\sigma})$ for any constant d. Then the least squares estimate based on $\widetilde{Y}_i$ is just $\widehat{\boldsymbol{\beta}}$, and the residuals are $d\psi(\widehat{e}_i/\widehat{\sigma})$.

Table 5 shows results for situations similar to Table 4 but with errors from t-transformed exponential variogram models. That is, we transformed the exponential errors used in Table 4 using the transformation $F_5^{-1}(\Phi(Y_i/\sqrt{3}))$, where Φ is the standard normal distribution function and F_5^{-1} is the inverse distribution function of a t distribution with 5 degrees of freedom. Thus the marginal distribution of Y_i is a t distribution with 5 degrees of freedom.

The entries in rows 1 and 2 and 5 and 6 of Table 5 are Monte Carlo estimates of the standard deviation of $\widehat{\boldsymbol{\beta}}$ based on 10,000 replications. These show that the robust regression

estimates have less variability than the least squares estimates as expected. The other rows are our standard deviation estimates for the regression parameter estimates divided by the estimated true standard deviations. The results are fairly similar to Table 4 and show that the estimates are negatively biased but less biased for the data with less correlation.

Splus version 3.4 along with the spatial module was used for all calculations.

5. Discussion

Our main goal has been to estimate the variance of regression coefficients obtained from a general estimating equations approach without making parametric semivariogram assumptions. Thus we have proposed a new method for non-parametric semivariogram estimation based on monotonizing a reduced-bias empirical semivariogram based on residuals.

The simulations of section 4 indicate that the resulting variance estimates are reasonably unbiased, and further simulations (not shown) illustrate convergence as the sample size grows. Comparison with REML using the true Matérn model and WNLS with the correct model shows that the new method is competitive with parametric methods. Table 4 also shows that our non-parametric approach compares even more favourably when misspecified parametric models are used. Our estimates have one large advantage not illustrated by these results: our approach is invariant to the presence or absence of a nugget effect. In contrast, with a nugget effect the Matérn models would need to be fitted with an added nugget parameter and results then are more variable.

Acknowledgements

We would like to thank John Coulston and USDA Forest Service (grant number SRS 00-CA-11330146-077) for providing the data of example 1 and for support during part of the research project. We are also grateful for helpful suggestions from the reviewers, one of whom in particular provided many detailed comments.

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–858.

Barlow, R., Bartholemew, D., Bremner, J. & Brunk, H. (1972). *Statistical inference under order restrictions*. Wiley, New York.

Barry, J. & Ver Hoef, J. (1996). Blackbox kriging: spatial prediction without specifying variogram models. *J. Agric. Biol. Environ. Statist.* **3**, 297–322.

Brownie, C. & Gumpertz, M. (1997). Validity of spatial analyses for large field trials. *J. Agric. Biol. Environ. Statist.* **2**, 1–23.

Cherry, S., Banfield, J. & Quimby, W. F. (1996). An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes. *J. Appl. Statist.* **23**, 435–449.

Cressie, N. (1987). A non-parametric view of generalized covariances for kriging. *Math. Geol.* **19**, 425–449.

Cressie, N. (1993). *Statistics for spatial data*. Wiley, New York.

Gorsich, D. J. & Genton, M. G. (2000). Variogram model selection via non-parametric derivative estimation. *Math. Geol.* **32**, 249–270.

Hall, P., Fisher, N. I. & Hoffman, B. (1994). On the non-parametric estimation of covariance functions. *Ann. Statist.* **22**, 2115–2134.

Huber, P. (1980). *Robust statistics*. Wiley, New York.

Jones, P. D., Parker, D. E., Osborn, T. J. & Briffa, K. R. (2000). Global and hemispheric temperature anomalies – land and marine instrumental records. In *Trends: a compendium of data on global change*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN. http://cdiac.esd.ornl.gov/trends/temp/jonescru/jones.html

Journel, A. & Huijbregts, C. (1979). *Mining geostatistics*. Academic Press, New York and London.

Koul, H. L. (1977). Behavior of robust estimators in the regression model with dependent errors. *Ann. Statist.* **5**, 681–699.

Lefohn, A. S. & Runeckles, V. C. (1987). Establishing a standard to protect vegetation – ozone exposure/dose considerations. *Atmos. Environ.* **21**, 561–568.

Lele, S. (1991). Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes. *J. Roy. Statist. Soc. Ser. B.* **53**, 253–267.

Lele, S. (1995). Inner product matrices, kriging, and non-parametric estimation of the variogram. *Math. Geol.* **27**, 673–692.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lumley, T. & Heagerty, P. J. (1999). Weighted empirical adaptive variance estimators for correlated data regression. *J. Roy. Statist. Soc. Ser. B.* **61**, 459–477.

Matérn, B. (1960). Spatial variation. *Meddelanden från Statens Skogsforskningsinstitut* **49**, No.5, Stockholm.

Matheron, G. (1973). The intrinsic random functions and their applications. *Adv. Appl. Probab.* **5**, 439–468.

Newey, W. K. & West, K. D. (1987). A simple positive semi-definite, heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.

Rousseeuw, P. J. & Molenberghs, G. (1993). Transformation of non-positive semidefinite correlation matrices. *Comm. Statist. Theory Methods* **22**, 965–984.

Shapiro, A. & Botha J. D. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Comput. Statist. Data Anal.* **11**, 87–96.

White, H. & Domowitz, I. (1984). Nonlinear regression with dependent observations. *Econometrica* **52**, 143–161.

Yasui, Y. & Lele, S. (1997). A regression method for spatial disease rates: an estimating function approach. *J. Amer. Statist. Assoc.* **92**, 21–32.

Received March 2002, in final form September 2003

Hyon-Jung Kim, Department of Mathematical Sciences and Statistics, PO Box 3000, University of Oulu, FIN-90014, Oulu, Finland.
E-mail: hyon-jung.kim@oulu.fi