Kriging with Nonparametric Variance Function Estimation

J. D. Opsomer,1* D. Ruppert,2 M. P. Wand,3 U. Holst,4 and O. Hössjer4
1Department of Statistics, Iowa State University, Ames, Iowa, U.S.A.
2School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York, U.S.A.
3Department of Biostatistics, Harvard University, Cambridge, Massachusetts U.S.A.
4Center for Mathematical Sciences, Division of Mathematical Statistics, University of Lund, Lund, Sweden, U.S.A.
*email: jopsomer@iastate.edu

SUMMARY. A method for fitting regression models to data that exhibit spatial correlation and heteroskedasticity is proposed. It is well known that ignoring a nonconstant variance does not bias least-squares estimates of regression parameters; thus, data analysts are easily lead to the false belief that moderate heteroskedasticity can generally be ignored. Unfortunately, ignoring nonconstant variance when fitting variograms can seriously bias estimated correlation functions. By modeling heteroskedasticity and standardizing by estimated standard deviations, our approach eliminates this bias in the correlations. A combination of parametric and nonparametric regression techniques is used to iteratively estimate the various components of the model. The approach is demonstrated on a large data set of predicted nitrogen runoff from agricultural lands in the Midwest and Northern Plains regions of the U.S.A. For this data set, the model comprises three main components: (1) the mean function, which includes farming practice variables, local soil and climate characteristics, and the nitrogen application treatment, is assumed to be linear in the parameters and is fitted by generalized least squares; (2) the variance function, which contains a local and a spatial component whose shapes are left unspecified, is estimated by local linear regression; and (3) the spatial correlation function is estimated by fitting a parametric variogram model to the standardized residuals, with the standardization adjusting the variogram for the presence of heteroskedasticity. The fitting of these three components is iterated until convergence. The model provides an improved fit to the data compared with a previous model that ignored the heteroskedasticity and the spatial correlation.

KEY WORDS: Heteroskedasticity; Local linear estimation; Metamodel; Runoff modeling; Spatial correlation.

1. Introduction

For many practical problems, the degree to which components of the statistical model can be specified in a parametric form varies dramatically. When the model is misspecified, the resulting model fit can be biased, and the possibility for making wrong inferences exists. However, when part of the model is amenable to parametric fitting, it is useful to do this to have a more analytically tractable model and to be able to use traditional inference techniques. Even in the most common form of nonparametric regression where the mean function is left unspecified, it is common to assume that the observations are uncorrelated, which can be viewed as a "parametric" assumption on the distribution of the errors. Violation of that assumption has a serious effect on the optimal bandwidth for estimating that mean function (e.g., Opsomer, 1997).

Most models for spatial data assume a stationary process that implies a constant variance. When the data are heteroskedastic, naively assuming a constant variance when fitting a variogram can lead to badly biased estimates of the correlation function. To appreciate this problem, one need only consider that the variance of the difference between the two observations depends not only on their correlation but also on their individual variances. In our experience, heteroskedasticity is common in spatial data but rarely can be fit by a parametric model.

In this article, we consider an application where it appears reasonable to accept a (roughly) linear relationship between dependent and independent variables and where the observations clearly display spatial dependence, but where the shape of the spatial variance cannot be specified a priori. The proposed approach blends elements of parametric and nonparametric fitting and is applicable to a wide range of problems, particularly those that entail spatially distributed observations.

We begin by describing the application that motivated this research. Economists at the Center for Agricultural and Rural Development at Iowa State University (CARD) are developing models to evaluate the impact of federal and state agricultural policies on the nitrogen water pollution in the Midwest and Northern Plains of the U.S.A. (Wu, Lakshminarayan, and Babcock, 1996), at both the regional and lo-
Nitrogen pollution occurs via two primary pathways: by nitrogen runoff into surface waters and by leaching through the soil into the groundwater. In the current article, we will focus on the prediction of nitrogen runoff. Table 1 shows the variables used in the model. They are further described in Wu et al. (1996). A map of the study region containing the locations of weather stations is given in Figure 1. The estimated variance function also displayed there will be discussed later.

Nitrogen runoff from nonpoint sources such as agricultural practices is typically unobservable, especially at the scale of interest in this study. The Water Quality and Erosion Productivity Impact Calculator (EPIC-WQ; see Sharpley and Williams, 1990), a widely used deterministic biogeophysical process model, provides, at least conceptually, a convenient tool for predicting the nitrogen runoff at the NRI points. Running the model for all NRI points would be computation intensive, and any change in any of the input variables would require rerunning the EPIC-WQ model. It was therefore decided to estimate a statistical “metamodel” on a representative subset of 11,403 data points and to use this metamodel in place of EPIC-WQ to predict the nitrogen runoff at the remaining observation points. Another advantage of this approach is the estimation of coefficients and the accompanying confidence intervals for the covariate effects, which provide additional insights into the nature of the effect of agricultural practices (represented by NRATE and the dummy variables in Table 1) on nitrogen pollution.

The original approach of Wu et al. (1996) was to fit the metamodel by ordinary least squares (OLS) after transforming the dependent variable and adding a limited number of interaction terms. The model was

\[
(Y N03)^{1/3.5} = \alpha + Z_1 \beta_{z_1} + \text{NRATE} * Z_2 \beta_{z_2} + X \beta_x + \text{i.i.d. errors},
\]

(1)

where \(Z_1\) contains the values for the covariates from Table 1 except the weather station location, \(Z_2\) is the same as \(Z_1\) except for the removal of the covariate NRATE, and \(X = \ldots\)
The data consist of \( n_i \) scalar response measurements \( Y_{ij} \) (the \( YNO3 \) measurements from Section 1) and covariates \( Z_{ij} \) recorded at \( N \) distinct geographic sites \( x_i \) (the weather stations from Section 1). The total number of observations is denoted by \( n = \sum_{i=1}^{N} n_i \).

The model is

\[
Y_{ij} = Z_{ij}' \beta + u_v(x_i)^{1/2} \xi_i + u_u(x_i)^{1/2} u_{ij} \tag{2}
\]

for \( j = 1, \ldots, n_i \) and \( i = 1, \ldots, N \). Here, \( \beta \) is a \( q \times 1 \) vector of parameters and \( u_v \) and \( u_u \) are bivariate variance functions. The errors \( u_{ij} \) are independent and identically distributed with \( E(u_{ij}) = 0 \) and \( \text{var}(u_{ij}) = 1 \). The \( \xi_i \) are such that \( E(\xi_i) = 0 \), \( \text{var}(\xi_i) = 1 \), and \( \text{cov}(\xi_i, \xi_j) = \rho(\|x_i - x_j\|; \theta) \), where \( \rho(\cdot; \theta) \) represents a parametric family of stationary, isotropic correlation functions indexed by the parameter \( \theta \).

The \( \{u_{ij}\} \) are independent of the \( \{\xi_i\} \), and both are independent of the \( \{Z_{ij}\} \). Model (2) is typical of variance components models where all within-site correlations are captured by the \( \{\xi_i\} \) so that the \( \{u_{ij}\} \) are independent. However, the \( \{\xi_i\} \) are modeled by a spatial process to allow between-site correlations.

This model is easily adapted to other situations. The mean function \( Z_{ij}' \beta \) can be replaced by any other parametric model, including \( Z_{ij}' \beta = \mu \) if ordinary kriging is used. Similarly, if there are no replicates at the geographic sites \( x_i \), \( i = 1 \) for all \( i \), the term \( u_u(x_i)^{1/2} u_{ij} \) can be subsumed into \( u_v(x_i)^{1/2} \xi_i \).

As mentioned earlier, many points share the same “location” \( x_i \), with \( n_i \) ranging from 1 to 221 for the \( N = 329 \) weather stations in our data set. There is also a computational reason for working with these approximate locations instead of the actual point locations: Only this reduction in the true dimension of the spatial variance–covariance matrix allows us to use “off-the-shelf” packages to perform the computations. The remaining errors \( u_{ij} \) at a given weather station location \( x_i \) were assumed to be independent, as the correlation is taken to be spatial. In the kriging context, the variance function associated with the \( u_{ij} \) is referred to as the nugget effect. If no replicates are available, the nugget effect would be estimated from the spatial error process \( v_u(x_i)^{1/2} \).

### 3. Estimation Procedure

#### 3.1 Overview

Let \( Y \) be the \( n \times 1 \) vector of \( Y_{ij} \)'s and \( Z \) be the \( n \times q \) matrix with the \( (i,j) \)th row equal to \( Z_{ij}' \). Let \( \Sigma \) be the variance–covariance matrix of \( Y \). Let \( p \) be a positive integer-valued tuning parameter. The role of \( p \) is to determine the minimum number of replicates needed at an \( x_i \) to use that location for estimating the variance functions. The choice of \( p \) is discussed later.

**Step 0.** (Initialization step) Set \( \bar{\Sigma} = I \).

**Step 1.** Obtain

\[
\hat{\beta} = (Z^T \bar{\Sigma}^{-1} Z)^{-1} Z^T \bar{\Sigma}^{-1} Y
\]

**Step 2.** Set

\[
r_{ij} = Y_{ij} - Z_{ij}^T \hat{\beta} \quad \text{and} \quad \bar{r}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}.
\]

**Step 3.** Obtain \( \hat{\nu}_v(x_i) \) by local linear smoothing of \( \{u_v(x_i); n_i \geq p\} \), where

\[
\hat{\nu}_v(x_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{r_{ij} - \bar{r}_i}{\bar{r}_i}.
\]

**Step 4.** Obtain \( \hat{\nu}_u(x_i) \) by local linear smoothing of \( \{u_u(x_i); n_i \geq p\} \), where

\[
\hat{\nu}_u(x_i) = \frac{\bar{r}_i}{n_i} - \frac{\hat{\nu}_v(x_i)}{n_i}.
\]

**Step 5.** Define \( \hat{v}_r(x_i) = \hat{\nu}_v(x_i) + \hat{\nu}_u(x_i) / n_i \) and let \( \bar{\varepsilon}_i = \bar{r}_i / \hat{v}_r(x_i)^{1/2} \). Estimate \( \theta \) in correlation model \( \rho(\cdot; \theta) \) by fitting the variogram of the \( \bar{\varepsilon}_i \).

**Step 6.** Obtain

\[
\bar{\Sigma} = \bar{\Sigma}_e + \bar{\Sigma}_u,
\]

where \( \{\bar{\Sigma}_u\}_{ij, ij'} = \hat{\nu}_u(x_i) \) if \( i = i' \), \( j = j' \) and 0 otherwise, and

\[
[\bar{\Sigma}]_{ij, ij'} = \hat{\nu}_v(x_i)^{1/2} \hat{\nu}_v(x_i')^{1/2} \rho(\|x_i - x_i'\|; \hat{\theta}).
\]
Step 7. Repeat steps 1–6 \( R_{\text{iter}} \) times.

Of course, the local linear smoothing in steps 3 and 4 could be replaced by higher-degree local polynomial regression. After the estimation steps have been completed, predictions can be made, as will be discussed in Section 5.

3.2 Details on the Implementation

Generalized least squares. In step 1, computations involving the inverse of the 11,403 \( \times \) 11,403 matrix \( \Sigma = \text{cov}(Y) \) are avoided by noting that, because of the assumed model (2),

\[
\Sigma = \Sigma_u + K^T V_e K,
\]

where \( \Sigma_u \) is a diagonal matrix with repeating “blocks” of length \( n_i; \)

\[
\Sigma_u = \text{block} \text{diag}\{v_u(x_i) I_{n_i} \ i = 1, \ldots, N\},
\]

with \( I_{n_i} \) being the \( n_i \times n_i \) identity matrix, \( V_e \) the \( N \times N \) covariance matrix of the \( \varepsilon_i \), and \( K \) an \( N \times n \) matrix with \( (i, j) \) entry equal to 1 for \( i = 1 + \sum_{k=1}^{j} n_k, \ldots, i = n \) and zero otherwise. The inverse of \( \Sigma \) is therefore equal to

\[
\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1} K^T (V_e^{-1} + K \Sigma_u^{-1} K^T)^{-1} K \Sigma_u^{-1},
\]

(Horn and Johnson, 1985, p. 19), which can be rapidly computed because the largest nondiagonal matrix to invert is only \( N \times N \).

Variance function estimation by local linear regression. The \( \hat{v}_u(x_i) \) in step 3 are approximately independently distributed, heteroskedastic random variables, with variance equal to \( 2v_u(x_i)^2/(n_i - 1) \). This would hold exactly if \( \beta \) were known and the errors were normally distributed. We will therefore apply the theory developed in Ruppert et al. (1997) to construct an estimator for the function \( v_u \). Although \( p = 2 \) observations are sufficient for computing \( \hat{v}_u(x_i) \) at a location, there is clearly much more information about \( v_u \) at locations with more observations. As \( n = 11,403 \), \( N = 329 \), and \( \bar{n} = 35 \), it might make sense to use only locations where \( n_i \) is “not too small.” We experimented with \( p = 2, 3, \) and 4 and found that \( p = 3 \) gave the best estimates, in terms of speed of convergence of the algorithm, and avoided boundary problems and negative variance estimates (see below). The number of locations where \( n_i \geq 3 \) is 290. The special structure between the estimator and its variance is used in the bandwidth selection of the EBBS algorithm (Ruppert, 1997). More specifically, let \( \hat{v}_u(x_i; h) \) be the local linear estimator of \( v_u(x_i) \). EBBS separately estimates the squared bias and variance of \( \hat{v}_u(x_i; h) \). These quantities are added together and their sum is minimized over a grid of \( h \) values to produce the EBBS bandwidth at \( x_i \). The bias estimate matches that in Ruppert (1997).

The estimate of \( \text{var}(\hat{v}_u(x_i; h)) \) uses the relation

\[
\text{var}(\hat{v}_u(x_i; h)) = s(x_i; h)^T \text{diag}\{\text{var}(\hat{v}_u(x_i))\} s(x_i; h),
\]

where \( s(x_i; h) \) is the \( N \) by 1 local polynomial “smoother vector” for a given value of \( h \), such that \( \hat{v}_u(x_i; h) = (\hat{v}_u(x_1), \ldots, \hat{v}_u(x_N)) s(x_i; h) \). EBBS estimates \( \text{var}(\hat{v}_u(x_i; h)) \) by

\[
\text{var}(\hat{v}_u(x_i; h)) = s(x_i; h)^T \text{diag}\{2(\hat{v}_u(x_i; h))/(n_i - 1)\} s(x_i; h).
\]

We let \( \hat{\Sigma}_{\text{EBBS}} \) denote the EBBS bandwidth and let \( \hat{v}_u(x_i; \hat{\Sigma}_{\text{EBBS}}) \) be denoted by \( \tilde{v}_u(x_i) \).

In step 4, we obtain \( \tilde{v}_e \) by smoothing \( \{\tilde{v}_e(x_i) : n_i > p\} \), again using EBBS to select the bandwidth. We will ignore the error in \( \beta \) so that \( \tilde{r}_{ij} = v_e(x_i)(1/2)\epsilon_i + v_u(x_i)(1/2)\epsilon_{ij} \) and therefore \( \tilde{r}_i = v_e(x_i)(1/2)\epsilon_i + v_u(x_i)(1/2)\tilde{u}_{i} \). Since \( \tilde{v}_u(x_i) \) is unbiased for \( v_u(x_i) \),

\[
E(\tilde{v}_u(x_i)) = E(\tilde{r}_i)^2 - v_u(x_i) = v_e(x_i).
\]

Therefore, when we smooth the \( \{\tilde{v}_e(x_i)\} \), there is no bias term involving \( v_u \), and EBBS will properly estimate the bias of our final estimate of \( v_e \). One might consider estimating \( v_e \) by smoothing the \( \{\tilde{r}_i\} \) and then subtracting off an estimate of \( v_u(x_i)/n_i \); however, in this case, the bandwidth optimal for smoothing the \( \{\tilde{r}_i\} \) will not be optimal for the final estimate of \( v_e \). The EBBS bandwidth for smoothing \( \tilde{v}_e(x_i) \) requires an estimate of \( \text{var}(\tilde{v}_e(x_i; h)) \). Estimation of this variance is based upon the following results.

Let \( H = Z(\Sigma^{-1} Z)^{-1} Z^T \Sigma^{-1} \) represent the "hat" matrix from the estimation of the mean. Let \( \kappa \) be the \( N \times n \) matrix, with \( (i, j) \) entry equal to 1/\( n_i \) for \( j = 1 + \sum_{k=1}^{i} n_k \), \( \ldots \), \( i - 1 \sum_{k=1}^{i} n_k \) and zero otherwise. Finally, let \( A \otimes B \) denote the elementwise product of equsized matrices \( A \) and \( B \).

Result 1. Assuming normality of the \( Y_{ij} \)'s, the covariance matrix containing \( \tilde{r}_i^2 \) for \( i = 1, \ldots, N \) is given by

\[
2A \Sigma A^T \otimes (A \Sigma A^T + 2A \kappa \kappa^T A^T),
\]

where \( m = Z\beta \) and \( A = \kappa(I - H) \). If the error due to estimation of the mean \( m \) is ignored, then the above covariance matrix simplifies to

\[
2(V_e + V_u E_1)^{[2]},
\]

where \( V_u = \text{diag}\{v_u(x_i)\} \) for \( i = 1, \ldots, N \), \( E_1 = \text{diag}\{1/n_i, i = 1, \ldots, N\}, \) and \( A^{[2]} = A \otimes A \).

Result 2. If the error due to estimation of \( \beta \) is ignored and the \( Y_{ij} \)'s are normally distributed, the covariance matrix of the random vector containing \( \hat{v}_e(x_i) = (\tilde{r}_i)^2 - \hat{v}_u(x_i)/n_i \), \( i = 1, \ldots, N \) is given by

\[
\Sigma_{\hat{v}_e} = 2\{V_e + V_u E_1^{[2]} + V_u^{[2]} E_1^{[2]} E_2^{[2]}\},
\]

where \( E_2 = \text{diag}\{(n_i - 1)^{-1}\} \).

Let \( \hat{\tilde{v}}_e \) and \( \hat{\tilde{v}}_u \) be obtained from \( V_e \) and \( V_u \) by replacing \( v_e(x_i) \) and \( v_u(x_i) \) by \( \hat{v}_e \) and \( \hat{v}_u \), respectively. Then let \( \Sigma_{\hat{v}_e} \) be given by (5) with \( V_e \) and \( V_u \) replaced by \( \hat{V}_e \) and \( \hat{V}_u \). Suppose that

\[
\hat{v}_e = (\hat{v}_e(x_1), \ldots, \hat{v}_e(x_N)) s(x_i; h),
\]

where, as with \( \hat{v}_u \), \( s(x_i; h) \) is a smoother vector. Then, the estimate of \( \text{var}(\hat{v}_e(x_i)) \) used by EBBS is

\[
s^T(x_i; h) \hat{\Sigma}_{\hat{v}_e} s(x_i; h).
\]

Our estimate of \( v_e \) does not use locations where \( n_i < p \), but these are locations where there is relatively little information about \( v_e \). Because the \( \tilde{v}_e(x_i) \) smoothed in step 4 are possibly negative, there is a positive probability that \( \tilde{v}_e(x_i) \) is negative.
As \( n_i \) increases, the probability that \( \hat{v}_e(x_i) \) is negative decreases. Although negative values for \( \hat{v}_e(x_i) \) are in principle not a problem, it is highly undesirable to have negative variance estimates \( v_e \), as they would result in a negative definite covariance matrix \( V_e \). For \( p = 3 \), only nine locations had negative variance estimates, and all were located at the north and north west boundaries of the estimation region, making it very likely that they are the result of “boundary effects,” a common nuisance in nonparametric regression similar to extrapolation problems in parametric regression. We therefore decided to add a local averaging step at each iteration of the algorithm to “correct” any negative estimates. Note that this step only changes the negative estimates and leaves all the other values unchanged.

**Estimation of the correlation function by variogram fitting.** In step 5, the correlation function is estimated parametrically by variogram fitting. Because heteroskedasticity is known to cause spurious patterns in variograms, it is important to remove that effect before estimating the correlation function. Hence, the spatial residuals \( \hat{r}_i \) have to be normalized. What the normalizing constants should be is a somewhat subtle issue. If we ignore the errors caused by the estimation of the mean and variance functions and use \( \hat{\sigma}_i = \hat{\sigma}_i / v_e(x) \), the variogram will estimate

\[
2\gamma(x_i - x_i') := E(\hat{\sigma}_i^2 - \hat{\sigma}_i'^2) = \text{var}(\hat{\sigma}_i) + \text{var}(\hat{\sigma}_i') - 2\text{cov}(\hat{\sigma}_i, \hat{\sigma}_i') = \frac{v_r(x_i)}{v_e(x_i)} + \frac{v_r(x_i')}{v_e(x_i')},
\]

whereas if we use \( \hat{\sigma}_i = \hat{\sigma}_i / v_r(x_i) \), then

\[
2\gamma'(x_i - x_i') := E(\hat{\sigma}_i^2 - \hat{\sigma}_i'^2) = 2 - 2\rho(x_i - x_i') \frac{v_r(x_i) v_r(x_i')}{v_r(x_i) v_r(x_i')},
\]

Neither \( \gamma(\cdot) \) nor \( \gamma'(\cdot) \) is generally equal to \( \gamma(\cdot) := 1 - \rho(\cdot) \), so they cannot be directly used to fit the correlation function. However, it is easy to see that

\[
\gamma(x_i - x_i') = 1 - \frac{1 - \gamma'(x_i - x_i')} {\frac{v_r(x_i)}{v_r(x_i)} + \frac{v_r(x_i')}{v_r(x_i')}}, \quad (6)
\]

We can therefore construct a “bias-corrected” variogram based on (6). Let \( \hat{\gamma}_i = \hat{\sigma}_i / v_r(x_i) \). For a given distance \( t \), let \( S(t) = \{(i, i') : ||x_i - x_i'|| \in (t \pm \delta)\} \) with \( \delta \) a given bin size and \( n(t) = |S(t)| \). The \( \delta \) was chosen so that 200 equal-sized bins were produced per the range of \( ||x_i - x_i'|| \) in the study region, corresponding to \( \delta \approx 0.09\text{°}. \) This represents a compromise between the computational tractability and the need for sufficient observations in each bin. Then,

\[
\hat{\gamma}(t) = 1 - \frac{1}{2n(t)} \sum_{S(t)} (\hat{\sigma}_i - \hat{\sigma}_i')^2
\]

The following parametric model is used for \( \rho(\cdot) \):

\[
\rho(t; \theta) = 1 - \theta_3 e^{-\theta_1 t} - (1 - \theta_3) e^{-\theta_2 t},
\]

with \( \theta_1, \theta_2 > 0 \) and \( 0 \leq \theta_3 \leq 1 \). This is a mixture of two exponential functions, which was chosen to guarantee the positive definiteness of the variance–covariance matrix estimate. Clearly, other parametric models, including mixtures of larger numbers of exponentials, could be selected as correlation functions for other data sets. The parameters \( \theta_1, \theta_2, \) and \( \theta_3 \) are estimated by weighted least-squares minimization following Cressie (1993, p. 96).

The estimate of the spatial variance–covariance matrix \( V_e \) is computed by setting

\[
[V_e]_{ii'} = \rho(x_i - x_i'; \hat{\theta}) \sqrt{v_e(x_i) v_e(x_i')},
\]

4. Results

The model was run on the CARD data set and converges in 2–10 iterations, depending on the strictness of the convergence criterion and on the choice of some of the tuning parameters. For \( p = 3 \), the model converges after four iterations, which takes approximately 10 minutes to run on a DEC 3000 Model 900 AXP workstation, with the bulk of the computing time taken by the generalized least-squares (GLS) fitting (step 1 in Section 3.1).

Figures 1 and 2 show the nonparametric estimates of the variance functions \( v_u(\cdot) \) and \( v_e(\cdot) \) at the weather station locations. Both estimates show a pattern of low values in the center. The estimated functions also display some interesting differences: The Great Lakes region exhibits high local and spatial variance, and the spatial variance is also high in the southernmost part of the study region, whereas the local variance is high at the western edge. Most of the variability in the data is explained by the local variance \( v_u \), with the

![Figure 2. Estimate of the variance function \( v_e(\cdot) \) at the weather station locations.](image-url)
mean value of $\hat{\varepsilon}_u(x_i)$ and $\hat{\varepsilon}_c(x_i)$ equal to 4.789 and 0.404, respectively.

In Figure 3, the bias-adjusted variogram of the standardized residuals $\tilde{\varepsilon}$ is displayed, as well as the weighted least-squares fitted variogram function. The spatial correlation decreases rapidly as distance increases and is only important for closely spaced points.

The goodness of fit of a model such as this can be evaluated using data-splitting techniques (e.g., Picard and Berk, 1990), and this approach was applied to a comparison between using the transformed and untransformed EPIC-WQ predicted nitrogen runoff values as dependent variables. This analysis was performed by fitting the model on 90% of the data and predicting on the remaining 10%, and it did not show any significant difference in average prediction error between the transformed and untransformed models. As mentioned in Section 1, this is not surprising because the heteroskedasticity is now explicitly accounted for in the model itself.

5. Model Predictions

The purpose of developing this metamodel is to facilitate the prediction of the potential nitrogen runoff at a set of 128,591 NRI points. As the prediction and estimation points use the same set of weather station locations, the spatial residuals $\varepsilon_i$ can be considered a lattice process (Cressie, 1993). The vector of spatial errors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$ can therefore be predicted by a “shrunk” version of the spatial residuals $\tilde{\varepsilon}$:

$$\tilde{\varepsilon} = \hat{V}_e (\hat{V}_e + \hat{V}_c E_1)^{-1} \tilde{r},$$

with $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_N)^T$ by a straightforward application of conditional expectations (e.g., Bickel and Doksum, 1977, p. 26). Hence, the spatial “correction” for an NRI point with closest weather station location $x_i^*$ can be predicted by the corresponding element of the vector $\tilde{\varepsilon}$. Figure 5 in Opsomer et al. (1997) shows a plot of the values of the spatial corrections $\tilde{\varepsilon}_i$.

6. Conclusions

We have described a method for fitting spatial data that combines parametrically specified mean and correlation functions with an unspecified spatial variance function. It can easily be generalized to other situations with different parametric models, or to situations without replication at the spatial locations. An iterative procedure for estimating the parameters and nonparametric regressions was explained in this article. However, no attempt was made to prove optimality or convergence properties for our algorithm, nor to more than sketch its theoretical properties under simplifying assumptions. These are topics for future research.

ACKNOWLEDGEMENTS

Research of Dr O. Hössjer was supported by the Swedish National Board for National and Technical Development Grant 91-02637P, by the National Science Foundation Grant DMS-9626762, and by the Center for Agricultural and Rural Development at Iowa State University.

The authors thank Wayne Fuller for his helpful suggestions in creating the bias-adjusted variogram.


**APPENDIX**

**Proof of Result 1**

The vector of residuals $r_{ij}$ can be written as $r = (I - H)Y$. From the definition of $\kappa$, we have the vector of $\bar{v}_i$ values equaling $\kappa(I - H)Y$. The stated result in (3) then follows directly from Lemma 1 of Ruppert et al. (1997) for the special case of normal $Y_i$.

If the bias due to estimation of $Z\beta$ is ignored, then (3) simplifies to

$$2(\kappa \Sigma_k^T)^{[2]}.$$

Expression (4) follows directly after noting that we are indexing the matrix $\Sigma$ as $\Sigma_{ij,j',j'} = \text{cov}(Y_{ij}, Y_{ij'})$ for $j = 1, \ldots, n_i, i = 1, \ldots N$. Hence,

$$(\kappa \Sigma_k^T_{ij}) = \begin{cases} \frac{1}{n_i n_j} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_j} \Sigma_{ij,j',j'}, & i = i', \\ \frac{\text{cov}(\bar{Y}_i)}{n_i}, & i \neq i', \\ \sqrt{\text{var}(\bar{Y}_i)\text{var}(\bar{Y}_{i'})\rho(||\bar{Y}_i - \bar{Y}_{i'}||)} & i \neq i'. \end{cases}$$

**Proof of Result 2**

Recall that

$$\hat{\nu}_i(x_i) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\bar{r}_{ij} - \bar{r}_i)^2 = \sum_{j=1}^{n_i} (\bar{u}_{ij} - \bar{u}_i)^2.$$ 

Because the $\{u_{ij}\}$ are i.i.d normals, $\bar{u}_i$ is independent of $\hat{\nu}_i(x_i)$. Therefore, $\hat{r}_i$ is independent of $\hat{\nu}_i(x_i)$. Let $\Sigma_r$ be the covariance matrix of the vector $(\hat{r}_{i1}, \ldots, \hat{r}_{iN})^T$, $\Sigma_{\hat{\nu}_i}$ the covariance matrix of $\{\hat{\nu}_i(x_1), \ldots, \hat{\nu}_i(x_N)\}$, and $\Sigma_{\hat{\nu}_i}$ the covariance matrix of $\{\hat{\nu}_i(x_1), \ldots, \hat{\nu}_i(x_N)\}^T$. As $\hat{\nu}_i(x_i)$ is $\text{var}(\bar{Y}_i)/(n_i - 1)$ times a $X^2(n_i - 1)$ random variable, we have

$$\Sigma_{\hat{\nu}_i} = \text{diag} \left(\frac{2\text{var}(\bar{Y}_i)}{n_i - 1}\right) = 2\Sigma_{\nu_i}[2].$$

By (4), and ignoring the error caused by using $\hat{\beta}$ in place of $\beta$, we have

$$\Sigma_{\hat{\nu}_i} = 2(\Sigma + \Sigma^{[2]}E_1)^{[2]} + \Sigma_{\nu_i}E_1^{[2]} = 2\{V_\epsilon + V_0E_1[2] + V_\nu^{[2]}E_1^{[2]}E_2[2]\}.$$  

$\square$