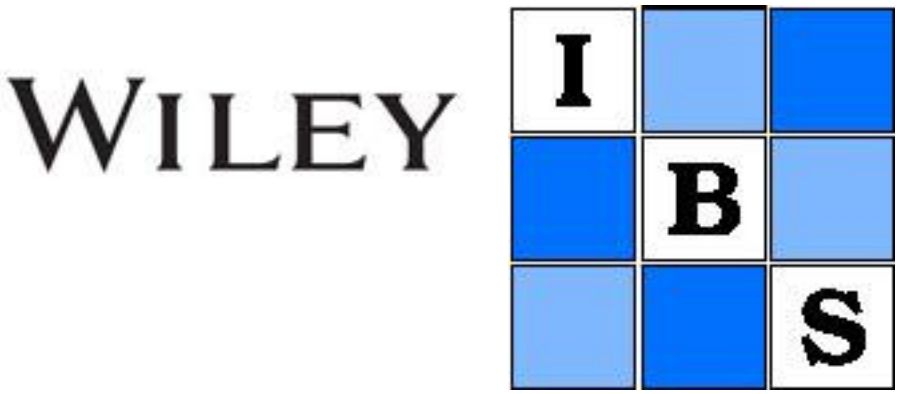

BIOMETRICS 55, 704–710
September 1999

Kriging with Nonparametric Variance Function Estimation

J. D. Opsomer,[1,*] D. Ruppert,[2] M. P. Wand,[3] U. Holst,[4] and O. Hössjer[4]

[1]Department of Statistics, Iowa State University, Ames, Iowa, U.S.A.
[2]School of Operations Research and Industrial Engineering,
Cornell University, Ithaca, New York, U.S.A.
[3]Department of Biostatistics, Harvard University, Cambridge, Massachusetts U.S.A.
[4]Center 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-

Table 1
Model variables

YN03	Nitrogen runoff (predicted by EPIC-WQ)
NRATE	Nitrogen application rate
	Tillage, Conservation, and Irrigation Practice Dummies (Reference: Conventional Tillage)
DRT	Reduced tillage
DSTRIP	Strip-cropping
DNT	No till
DTERRA	Terracing
DCONTR	Contouring
DIRTYP	Irrigation
	Crop Rotation Dummies (Reference: Continuous Alfalfa)
DROT1	Continuous corn
DROT8	Soybeans-soybeans-corn
DROT2	Continuous soybeans
DROT9	Wheat-fallow
DROT3	Continuous wheat
DROT10	Wheat-sorghun-fallow
DROT4	Continuous sorghum
DROT11	Wheat-soybeans
DROT5	Corn-soybeans
DROT12	Wheat-sorghum
DROT6	Corn-corn-soybeans
DROT14	Corn-corn-3 years alfalfa
DROT7	Corn-soybeans-wheat
	Rainfall and Soil Properties
RAIN	Rainfall (mm)
BD	Bulk density
SLOPE	Field slope
PH	Soil pH
CLAY	Clay percentage
PERM	Soil permeability
OM	Organic matter (%)
AWC	Available water capacity
	Hydrology Dummies (Reference: DHYGA)
DHYGB	Hydrologic group B
DHYGD	Hydrologic group D
DHYGC	Hydrologic group C
	Location of Closest Weather Station
LAT	Latitude
LONG	Longitude

cal levels. Local prediction is achieved by using the 128,591 National Resources Inventory (NRI) points in the region of interest as the basis for the evaluation of pollution impact: The NRI database provides measurements on many land-use and soil variables of interest, as well as sampling weights that allow statistically valid area predictions based on the point predictions (Nusser and Goebel, 1997).

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.

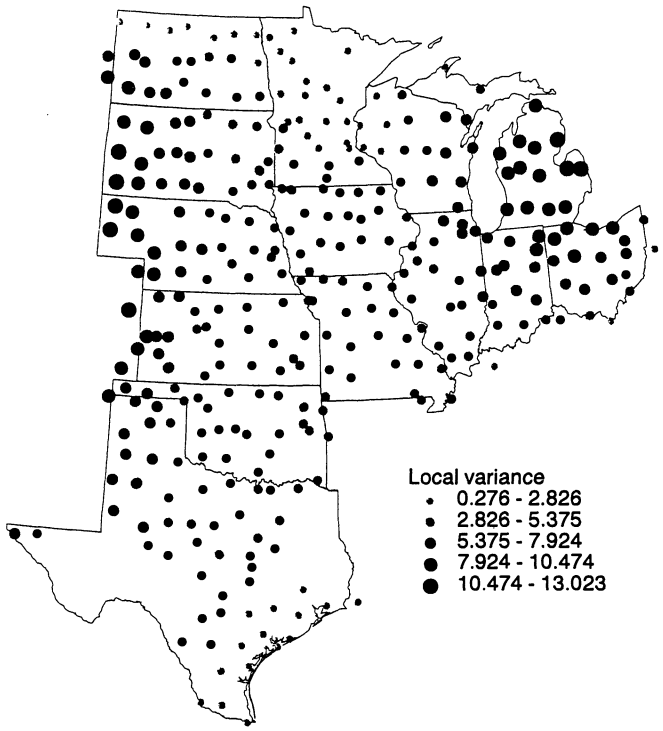

Figure 1. Estimate of the variance function $v_u(\cdot)$ at the weather station locations.

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

$$(YN03)^{1/3.5} = \alpha + \boldsymbol{Z}_1\boldsymbol{\beta}_{z_1} + \text{NRATE} * \boldsymbol{Z}_2\boldsymbol{\beta}_{z_2} + \boldsymbol{X}\boldsymbol{\beta}_x + \text{i.i.d. errors}, \quad (1)$$

where $\boldsymbol{Z}_1$ contains the values for the covariates from Table 1 except the weather station location, $\boldsymbol{Z}_2$ is the same as $\boldsymbol{Z}_1$ except for the removal of the covariate NRATE, and $\boldsymbol{X}$ =

(LAT, LONG), the location of the nearest weather station. We will let $\boldsymbol{Z} = [\boldsymbol{Z}_1 \text{ NRATE} * \boldsymbol{Z}_2 \ \boldsymbol{X}]$. For simplicity, we refer to $\boldsymbol{Z}$ as the covariates for this model and let $\boldsymbol{\beta} = [\boldsymbol{\beta}_{z_1}^{\mathrm{T}} \ \boldsymbol{\beta}_{z_2}^{\mathrm{T}} \ \boldsymbol{\beta}_x^{\mathrm{T}}]^{\mathrm{T}}$. The location and interaction terms were included to improve the fit of the model, and the transformation was selected to remove some of the observed departures from usual assumptions that errors are homoskedastic and normally distributed. Nevertheless, the residuals still exhibited severe heteroskedasticity, as well as spatial correlation. As noted in Carroll and Ruppert (1988), transformations of the dependent variable only remove heteroskedasticity when it depends on the mean. They are therefore not appropriate in cases where spatial location appears to cause most of the variance effects. This was confirmed for this data set: When using the proposed model, transformation of the dependent variable no longer had any noticeable effect on the goodness of fit of the model (see Section 4).

In the current paper, we demonstrate how a combination of universal kriging and nonparametric variance function estimation can be used to develop an improved regression model for this problem and simultaneously maintain the interpretability of mean function model (1). The choice of kriging is motivated by the fact that one of the primary uses of this model is the prediction of $YN03$ at a large number of points not included in the regression observations, a situation for which kriging has well-known optimality properties (Cressie, 1993). As the residuals of the OLS fit of model (1) exhibited significant heteroskedasticity as well, the explicit inclusion of a spatial variance function is expected to further improve the fit of both the mean and the correlation function. A generalization of the nonparametric variance estimation approach of Ruppert et al. (1997) is used to estimate the variance function.

Section 2 proposes a model that explicitly accounts for the heteroskedasticity and spatial correlation in the data, and Section 3 describes the approach used in estimating its various components. In Section 4, the model estimates are discussed. Section 5 addresses the use of universal kriging for predicting the nitrogen runoff values at the remaining NRI points not included in the metamodel.

2. The Model

The data consist of n_i scalar response measurements Y_{ij} (the $YN03$ measurements from Section 1) and covariates $\boldsymbol{Z}_{ij}$ recorded at N distinct geographic sites $\boldsymbol{x}_i$ (the weather stations from Section 1). The total number of observations is denoted by $n = \Sigma_{i=1}^{N} n_i$.

The model is

$$Y_{ij} = \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + v_\varepsilon(\boldsymbol{x}_i)^{1/2}\varepsilon_i + v_u(\boldsymbol{x}_i)^{1/2}u_{ij} \qquad (2)$$

for $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. Here, $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters and v_ε and v_u are bivariate variance functions. The errors u_{ij} are independent and identically distributed with $E(u_{ij}) = 0$ and $\mathrm{var}(u_{ij}) = 1$. The ε_i are such that $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = 1$, and $\mathrm{cov}(\varepsilon_i, \varepsilon_{i'}) = \rho(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|; \boldsymbol{\theta})$, where $\rho(\cdot; \boldsymbol{\theta})$ represents a parametric family of stationary, isotropic correlation functions indexed by the parameter $\boldsymbol{\theta}$. The $\{u_{ij}\}$ are independent of the $\{\varepsilon_i\}$, and both are independent of the $\{\boldsymbol{Z}_{ij}\}$. Model (2) is typical of variance components models where all within-site correlations are captured by the $\{\varepsilon_i\}$ so that the $\{u_{ij}\}$ are independent. However, the $\{\varepsilon_i\}$ are modeled by a spatial process to allow between-site correlations.

This model is easily adapted to other situations. The mean function $\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$ can be replaced by any other parametric model, including $\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} \equiv \mu$ if ordinary kriging is used. Similarly, if there are no replicates at the geographic sites $\boldsymbol{x}_i$ (i.e., $n_i = 1$ for all i), the term $v_u(\boldsymbol{x}_i)^{1/2}u_{ij}$ can be subsumed into $v_\varepsilon(\boldsymbol{x}_i)^{1/2}\varepsilon_i$.

As mentioned earlier, many points share the same "location" $\boldsymbol{x}_i$, 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 $\boldsymbol{x}_i$ 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 $v_\varepsilon(\boldsymbol{x}_i)^{1/2}\varepsilon_i$.

3. Estimation Procedure

3.1 *Overview*

Let $\boldsymbol{Y}$ be the $n \times 1$ vector of Y_{ij}'s and $\boldsymbol{Z}$ be the $n \times q$ matrix with the (i,j)th row equal to $\boldsymbol{Z}_{ij}^{\mathrm{T}}$. Let $\boldsymbol{\Sigma}$ be the variance–covariance matrix of $\boldsymbol{Y}$. Let p be a positive integer-valued tuning parameter. The role of p is to determine the minimum number of replicates needed at an $\boldsymbol{x}_i$ to use that location for estimating the variance functions. The choice of p is discussed later.

Step 0. (Initialization step) Set $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{I}$.

Step 1. Obtain

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}.$$

Step 2. Set

$$r_{ij} = Y_{ij} - \boldsymbol{Z}_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \quad \text{and} \quad \overline{r}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}.$$

Step 3. Obtain $\widehat{v}_u(\boldsymbol{x}_i)$ by local linear smoothing of $\{\widetilde{v}_u(\boldsymbol{x}_i)\colon n_i \geq p\}$, where

$$\widetilde{v}_u(\boldsymbol{x}_i) = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(r_{ij} - \overline{r}_i)^2.$$

Step 4. Obtain $\widehat{v}_\varepsilon(\boldsymbol{x}_i)$ by local linear smoothing of $\{\tilde{v}_\varepsilon(\boldsymbol{x}_i)\colon n_i \geq p\}$, where

$$\tilde{v}_\varepsilon(\boldsymbol{x}_i) = (\overline{r}_i)^2 - \frac{\tilde{v}_u(\boldsymbol{x}_i)}{n_i}.$$

Step 5. Define $\widehat{v}_r(\boldsymbol{x}_i) = \widehat{v}_\varepsilon(\boldsymbol{x}_i) + \widehat{v}_u(\boldsymbol{x}_i)/n_i$ and let $\widehat{\varepsilon}_i = \overline{r}_i/\widehat{v}_r(\boldsymbol{x}_i)^{1/2}$. Estimate $\boldsymbol{\theta}$ in correlation model $\rho(\cdot; \boldsymbol{\theta})$ by fitting the variogram of the $\widehat{\varepsilon}_i$.

Step 6. Obtain

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_\varepsilon + \widehat{\boldsymbol{\Sigma}}_u,$$

where $(\widehat{\boldsymbol{\Sigma}}_u)_{ij,i'j'} = \widehat{v}_u(\boldsymbol{x}_i)$ if $i = i'$, $j = j'$ and 0 otherwise, and

$$[\widehat{\boldsymbol{\Sigma}}_\varepsilon]_{ij,i'j'} = \widehat{v}_\varepsilon(\boldsymbol{x}_i)^{1/2}\widehat{v}_\varepsilon(\boldsymbol{x}_i')^{1/2}\rho(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|; \widehat{\boldsymbol{\theta}}).$$

Step 7. Repeat steps 1–6 R_{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 × 11,403 matrix $\widehat{\boldsymbol{\Sigma}} = \widehat{\text{cov}}(\boldsymbol{Y})$ are avoided by noting that, because of the assumed model (2),

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_u + \boldsymbol{K}^{\mathrm{T}}\boldsymbol{V}_\varepsilon\boldsymbol{K},$$

where $\boldsymbol{\Sigma}_u$ is a diagonal matrix with repeating "blocks" of length n_i:

$$\boldsymbol{\Sigma}_u = \text{block diag}\{v_u(\boldsymbol{x}_i)\, I_{n_i}\ i = 1, \ldots, N\},$$

with I_{n_i} being the $n_i \times n_i$ identity matrix, $\boldsymbol{V}_\varepsilon$ the $N \times N$ covariance matrix of the ε_i, and $\boldsymbol{K}$ an $N \times n$ matrix with (i, i') entry equal to 1 for $i' = 1 + \Sigma_{k=1}^{i-1} n_k, \ldots, \Sigma_{k=1}^{i} n_k$ and zero otherwise. The inverse of $\boldsymbol{\Sigma}$ is therefore equal to

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\boldsymbol{K}^{\mathrm{T}}\left(\boldsymbol{V}_\varepsilon^{-1} + \boldsymbol{K}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{K}^{\mathrm{T}}\right)^{-1}\boldsymbol{K}\boldsymbol{\Sigma}_u^{-1},$$

(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 $\tilde{v}_u(\boldsymbol{x}_i)$ in step 3 are approximately independently distributed, heteroskedastic random variables, with variance equal to $2v_u(\boldsymbol{x}_i)^2/(n_i - 1)$. This would hold exactly if $\boldsymbol{\beta}$ 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 $\tilde{v}_u(\boldsymbol{x}_i)$ at a location, there is clearly much more information about v_u at locations with more observations. As n = 11,403, N = 329, and $\overline{n}$ = 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 $\widehat{v}_u(\boldsymbol{x}_i; h)$ be the local linear estimator of $v_u(\boldsymbol{x}_i)$. EBBS separately estimates the squared bias and variance of $\widehat{v}_u(\boldsymbol{x}_i; h)$. These quantities are added together and their sum is minimized over a grid of h values to produce the EBBS bandwidth at $\boldsymbol{x}_i$. The bias estimate matches that in Ruppert (1997).

The estimate of $\text{var}(\widehat{v}_u(\boldsymbol{x}_i; h))$ uses the relation

$$\text{var}(\widehat{v}_u(\boldsymbol{x}_i; h)) = \boldsymbol{s}(\boldsymbol{x}_i; h)^{\mathrm{T}}\text{diag}\{\text{var}(\tilde{v}_u(\boldsymbol{x}_i)\}\boldsymbol{s}(\boldsymbol{x}_i; h),$$

where $\boldsymbol{s}(\boldsymbol{x}_i; h)$ is the N by 1 local polynomial "smoother vector" for a given value of h, such that $\widehat{v}_u(\boldsymbol{x}_i; h) = (\tilde{v}_u(\boldsymbol{x}_1), \ldots, \tilde{v}_u(\boldsymbol{x}_N))\boldsymbol{s}(\boldsymbol{x}_i; h)$. EBBS estimates $\text{var}(\widehat{v}_u(\boldsymbol{x}_i; h))$ by

$$\widehat{\text{var}}(\widehat{v}_u(\boldsymbol{x}_i; h)) = \boldsymbol{s}(\boldsymbol{x}_i; h)^{\mathrm{T}}\text{diag}\{2(\widehat{v}_u(\boldsymbol{x}_i; h))/(n_i - 1)\}\boldsymbol{s}(\boldsymbol{x}_i; h).$$

We let $\widehat{h}_{\text{EBBS}}$ denote the EBBS bandwidth and let $\widehat{v}_u(\boldsymbol{x}_i; \widehat{h}_{\text{EBBS}})$ be denoted by $\widehat{v}_u(\boldsymbol{x}_i)$.

In step 4, we obtain $\widehat{v}_\varepsilon$ by smoothing $\{\tilde{v}_\varepsilon(\boldsymbol{x}_i) : n_i > p\}$, again using EBBS to select the bandwidth. We will ignore the error in $\widehat{\boldsymbol{\beta}}$ so that $r_{ij} = v_\varepsilon(\boldsymbol{x}_i)(1/2)\epsilon_i + v_u(\boldsymbol{x}_i)(1/2)u_{ij}$ and therefore $\overline{r}_i = v_\varepsilon(\boldsymbol{x}_i)(1/2)\epsilon_i + v_u(\boldsymbol{x}_i)(1/2)\overline{u}_i$. Since $\tilde{v}_u(\boldsymbol{x}_i)$ is unbiased for $v_u(\boldsymbol{x}_i)$,

$$E(\tilde{v}_\varepsilon(\boldsymbol{x}_i)) = E(\overline{r}_i)^2 - \frac{v_u(\boldsymbol{x}_i)}{n_i} = v_\varepsilon(\boldsymbol{x}_i).$$

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

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

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

$$2\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}} \odot \left(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}} + 2\boldsymbol{A}\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right), \tag{3}$$

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

$$2(\boldsymbol{V}_\varepsilon + \boldsymbol{V}_u\boldsymbol{E}_1)^{[2]}, \tag{4}$$

where $\boldsymbol{V}_u = \text{diag}\{v_u(\boldsymbol{x}_i)$ for $i = 1, \ldots, N\}$, $\boldsymbol{E}_1 = \text{diag}\{1/n_i, i = 1, \ldots, N\}$, and $\boldsymbol{A}^{[2]} = \boldsymbol{A} \odot \boldsymbol{A}$.

Result 2. If the error due to estimation of $\boldsymbol{\beta}$ is ignored and the Y_{ij}'s are normally distributed, the covariance matrix of the random vector containing

$$\tilde{v}_\varepsilon(\boldsymbol{x}_i) = (\overline{r}_i)^2 - \frac{\tilde{v}_u(\boldsymbol{x}_i)}{n_i}, \quad i = 1, \ldots, N$$

is given by

$$\boldsymbol{\Sigma}_{\tilde{v}_\varepsilon} = 2\left\{(\boldsymbol{V}_\epsilon + \boldsymbol{V}_u\boldsymbol{E}_1)^{[2]} + \boldsymbol{V}_u^{[2]}\boldsymbol{E}_1^{[2]}\boldsymbol{E}_2\right\}, \tag{5}$$

where $\boldsymbol{E}_2 = \text{diag}((n_i - 1)^{-1})$.

Let $\widehat{\boldsymbol{V}}_\varepsilon$ and $\widehat{\boldsymbol{V}}_u$ be obtained from $\boldsymbol{V}_\varepsilon$ and $\boldsymbol{V}_u$ by replacing $v_u(\boldsymbol{x}_i)$ and $v_\varepsilon(\boldsymbol{x}_i)$ by $\widehat{v}_u$ and $\widehat{v}_e$, respectively. Then let $\widehat{\boldsymbol{\Sigma}}_{\tilde{v}_\varepsilon}$ be given by (5) with $\boldsymbol{V}_\varepsilon$ and $\boldsymbol{V}_u$ replaced by $\widehat{\boldsymbol{V}}_\varepsilon$ and $\widehat{\boldsymbol{V}}_u$. Suppose that

$$\widehat{v}_e = (\tilde{v}_\varepsilon(\boldsymbol{x}_1, \ldots, \tilde{v}_\varepsilon(\boldsymbol{x}_n))\boldsymbol{s}(\boldsymbol{x}_i; h),$$

where, as with $\widehat{v}_u$, $\boldsymbol{s}(\boldsymbol{x}_i; h)$ is a smoother vector. Then, the estimate of $\text{var}(\widehat{v}_\varepsilon(\boldsymbol{x}_i))$ used by EBBS is

$$\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{x}_i; h)\widehat{\boldsymbol{\Sigma}}_{\tilde{v}_\varepsilon}\boldsymbol{s}(\boldsymbol{x}_i; h).$$

Our estimate of v_ε does not use locations where $n_i < p$, but these are locations where there is relatively little information about v_ε. Because the $\tilde{v}_\varepsilon(\boldsymbol{x}_i)$ smoothed in step 4 are possibly negative, there is a positive probability that $\widehat{v}_\varepsilon(\boldsymbol{x}_i)$ is negative.

As n_i increases, the probability that $\tilde{v}_\varepsilon(\boldsymbol{x}_i)$ is negative decreases. Although negative values for $\tilde{v}_\varepsilon(\boldsymbol{x}_i)$ are in principle not a problem, it is highly undesirable to have negative variance estimates $\widehat{v}_e$, as they would result in a negative definite covariance matrix $\widehat{\boldsymbol{V}}_\varepsilon$. 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 $\overline{r}_i$ 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 $\tilde{\varepsilon}_i^* = \overline{r}_i / v_\varepsilon(\boldsymbol{x})^{1/2}$, the variogram will estimate

$$\begin{aligned} 2\tilde{\gamma}(\boldsymbol{x}_i - \boldsymbol{x}_{i'}) &:= E(\tilde{\varepsilon}_i^* - \tilde{\varepsilon}_{i'}^*)^2 \\ &= \operatorname{var}(\tilde{\varepsilon}_i^*) + \operatorname{var}(\tilde{\varepsilon}_{i'}^*) - 2\operatorname{cov}(\tilde{\varepsilon}_i^*, \tilde{\varepsilon}_{i'}^*) \\ &= \frac{v_r(\boldsymbol{x}_i)}{v_\varepsilon(\boldsymbol{x}_i)} + \frac{v_r(\boldsymbol{x}_{i'})}{v_\varepsilon(\boldsymbol{x}_{i'})} - 2\rho(\boldsymbol{x}_i - \boldsymbol{x}_{i'}), \end{aligned}$$

whereas if we use $\widehat{\varepsilon}_i^* = \overline{r}_i / v_r(\boldsymbol{x}_i)^{1/2}$, then

$$\begin{aligned} 2\tilde{\gamma}'(\boldsymbol{x}_i - \boldsymbol{x}_{i'}) &:= E(\widehat{\varepsilon}_i^* - \widehat{\varepsilon}_{i'}^*)^2 \\ &= 2 - 2\rho(\boldsymbol{x}_i - \boldsymbol{x}_{i'}) \sqrt{\frac{v_\varepsilon(\boldsymbol{x}_i)}{v_r(\boldsymbol{x}_i)} \frac{v_\varepsilon(\boldsymbol{x}_{i'})}{v_r(\boldsymbol{x}_{i'})}}. \end{aligned}$$

Neither $\tilde{\gamma}(\cdot)$ nor $\tilde{\gamma}'(\cdot)$ 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

$$\gamma(\boldsymbol{x}_i - \boldsymbol{x}_{i'}) = 1 - \frac{1 - \tilde{\gamma}'(\boldsymbol{x}_i - \boldsymbol{x}_{i'})}{\sqrt{\dfrac{v_\varepsilon(\boldsymbol{x}_i)}{v_r(\boldsymbol{x}_i)} \dfrac{v_\varepsilon(\boldsymbol{x}_{i'})}{v_r(\boldsymbol{x}_{i'})}}}. \tag{6}$$

We can therefore construct a "bias-corrected" variogram based on (6). Let $\widehat{\varepsilon}_i = \overline{r}_i / v_r(\boldsymbol{x}_i)^{1/2}$. For a given distance t, let $S(t) = \{(i, i') : \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| \in (t \pm \delta)\}$ with δ a given bin size and $n(t) = |S(t)|$. The δ was chosen so that 200 equal-sized bins were produced over the range of $\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|$ in the study region, corresponding to $\delta \approx 0.09°$. This represents a compromise between the computational tractability and the need for sufficient observations in each bin. Then,

$$\widehat{\gamma}(t) = 1 - \frac{1 - \dfrac{1}{2n(t)} \displaystyle\sum_{S(t)} (\widehat{\varepsilon}_i - \widehat{\varepsilon}_{i'})^2}{\dfrac{1}{n(t)} \displaystyle\sum_{S(t)} \sqrt{\dfrac{\widehat{\varepsilon}_r(\boldsymbol{x}_i)}{\widehat{v}_r(\boldsymbol{x}_i)} \dfrac{\widehat{v}_\varepsilon(\boldsymbol{x}_{i'})}{\widehat{v}_r(\boldsymbol{x}_{i'})}}}.$$

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

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

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 θ_1, θ_2, and θ_3 are estimated by weighted least-squares minimization following Cressie (1993, p. 96).

The estimate of the spatial variance–covariance matrix $\boldsymbol{V}_\varepsilon$ is computed by setting

$$[\widehat{\boldsymbol{V}}_\varepsilon]_{ii'} = \rho(\boldsymbol{x}_i - \boldsymbol{x}_{i'}; \widehat{\boldsymbol{\theta}}) \sqrt{\widehat{v}_\varepsilon(\boldsymbol{x}_i) \widehat{v}_\varepsilon(\boldsymbol{x}_{i'})}.$$

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 $v_\varepsilon(\cdot)$ 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

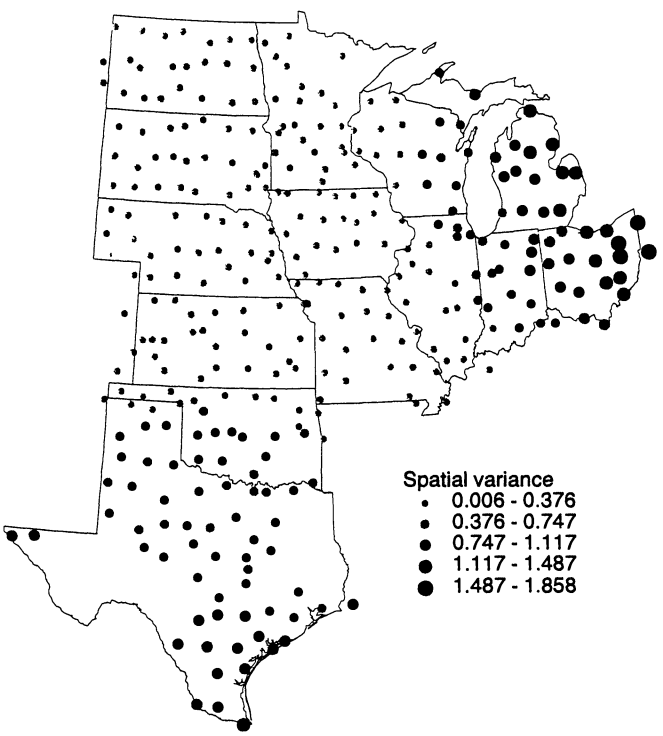

Figure 2. Estimate of the variance function $v_\varepsilon(\cdot)$ at the weather station locations.

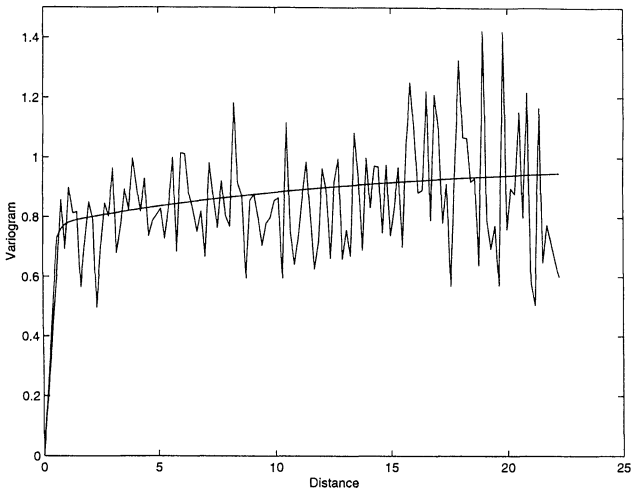

Figure 3. Variogram and estimated variogram function $\widehat{\rho}(\cdot;\widehat{\theta})$.

mean value of $\widehat{v}_u(\boldsymbol{x}_i)$ and $\widehat{v}_\varepsilon(\boldsymbol{x}_i)$ equal to 4.789 and 0.404, respectively.

In Figure 3, the bias-adjusted variogram of the standardized residuals $\widehat{\varepsilon}_i$ 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 ε_i can be considered a *lattice* process (Cressie, 1993). The vector of spatial errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^{\mathrm{T}}$ can therefore be predicted by a "shrunk" version of the spatial residuals $\bar{r}_i$:

$$\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{V}}_\varepsilon \left(\widehat{\boldsymbol{V}}_\varepsilon + \widehat{\boldsymbol{V}}_u \boldsymbol{E}_1\right)^{-1} \bar{\boldsymbol{r}}, \tag{7}$$

with $\bar{\boldsymbol{r}} = (\bar{r}_1, \ldots, \bar{r}_N)^{\mathrm{T}}$ 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 $\boldsymbol{x}_i^*$ can be predicted by the corresponding element of the vector $\widehat{\boldsymbol{\varepsilon}}$. Figure 5 in Opsomer et al. (1997) shows a plot of the values of the spatial corrections $\widehat{\varepsilon}_i$.

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.

Résumé

Nous présentons une méthode d'ajustement de modèles de régression à des données présentant une corrélation spatiale et une hétéroscédasticité. Il est bien connu qu'ignorer que la variance n'est pas constante ne biaise pas les estimations des moindres carrés des paramètres de la régression, si bien que ceux qui analysent de telles données sont facilement conduits à penser qu'une légère hétéroscédasticité peut être ignorée, ce qui est faux. Car, en l'ignorant l'ajustement des variogrammes conduit à des fonctions de corrélation biaisées. En modélisant l'hétéroscédasticité et en standardisant par l'écart type estimé, notre approche élimine ce biais dans les corrélations. Nous utilisons une combinaison de techniques de régression paramétrique et nonparamétrique pour estimer de manière itérative les différentes composantes du modèle. Nous démontrons cette approche sur un corpus de données important d'écoulement de l'azote prédit dans des terres agricoles du centre-ouest et du nord des États-Unis. Pour ce corpus, le modèle se sépare en trois composantes: (1) la fonction moyenne qui comprend les variables de la pratique agricole, les caractéristiques locales du sol et du climat et l'application du traitement azoté qui est supposée linéaire en ses paramètres et qui est ajustée par les moindres carrés généralisés, (2) la fonction de variance, qui contient une composante locale et une autre spatiale, dont la forme n'est pas spécifiée, est ajustée par un régression linéaire locale, et (3) la fonction de corrélation spatiale qui est estimée par l'ajustement d'un modèle de variogramme paramétrique aux résidus standardisés, avec la standardisation ajustant le variogramme en présence d'hétéroscédasticité. L'ajustement de ces trois composantes est effectué de manière itérative jusqu'à la convergence. Le modèle fournit un ajustement aux données amélioré si on le compare au modèle précédent qui ignorait hétéroscédasticité et corrélation spatiale.

References

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics.* Oakland, California: Holden-Day.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* New York: Chapman and Hall.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, 2nd edition. New York: Wiley.

Horn, R. A. and Johnson, C. A. (1985). *Matrix Analysis.* Cambridge, U.K.: Cambridge University Press.

Nusser, S. M. and Goebel, J. J. (1997). The national resources inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics* **4,** 181–204.

Opsomer, J.-D. (1997). Nonparametric regression in the presence of correlated errors. In *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, and R. D. Wolfinger (eds), 339–348. New York: Springer-Verlag.

Opsomer, J.-D., Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997). An application of kriging with nonparametric variance function estimation. In *1997 Proceedings of the Biometrics Section*, 123–128. Alexandria, Virginia: American Statistical Association.

Picard, R. R. and Berk, K. N. (1990). Data splitting. *American Statistician* **44,** 140–147.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* **92,** 1049–1062.

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39,** 262–273.

Sharpley, A. N. and Williams, J. R. (eds). (1990). *Epic-erosion/productivity impact calculator: 1. Model documentation.* Technical Bulletin 1768, USDA, Washington, D.C.

Wu, J., Lakshminarayan, P. G., and Babcock, B. A. (1996). Impacts of agricultural practices and policies on potential nitrate water pollution in the Midwest and the Northern Plains of the United States. Working Paper 96–WP 148, Center for Agricultural and Rural Development, Iowa State University, February.

Received January 1998. *Revised September* 1998. *Accepted October* 1998.

Appendix

Proof of Result 1

The vector of residuals r_{ij} can be written as $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. From the definition of $\boldsymbol{\kappa}$, we have the vector of $\overline{r}_i$ values equaling $\boldsymbol{\kappa}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. 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 $\boldsymbol{Z\beta}$ is ignored, then (3) simplifies to

$$2\left(\boldsymbol{\kappa\Sigma\kappa}^{\mathrm{T}}\right)^{[2]}.$$

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

$$\begin{aligned}\left(\boldsymbol{\kappa\Sigma\kappa}^{\mathrm{T}}\right)_{ii'} &= \frac{1}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} \Sigma_{ij,i'j'} \\ &= \begin{cases} \dfrac{v_u(\boldsymbol{x}_i)}{n_i} + v_\varepsilon(\boldsymbol{x}_i), & i = i', \\ \sqrt{v_\varepsilon(\boldsymbol{x}_i) v_\varepsilon(\boldsymbol{x}_{i'})}\,\rho(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|; \theta), & i \neq i', \end{cases}\end{aligned}$$

□

Proof of Result 2

Recall that

$$\tilde{v}_u(\boldsymbol{x}_i) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (r_{ij} - \overline{r}_i)^2 = \sum_{j=1}^{n_i} (u_{ij} - \overline{u}_i)^2.$$

Because the $\{u_{ij}\}$ are i.i.d. normals, $\overline{u}_i$ is independent of $\tilde{v}_u(\boldsymbol{x}_i)$. Therefore, $\overline{r}_i$ is independent of $\tilde{v}_u(\boldsymbol{x}_i)$. Let $\boldsymbol{\Sigma}_r$ be the covariance matrix of the vector $((\overline{r}_{1\cdot})^2, \ldots, (\overline{r}_{N\cdot})^2)^{\mathrm{T}}$, $\boldsymbol{\Sigma}_{\tilde{v}_\varepsilon}$ the covariance matrix of $(\tilde{v}_\varepsilon(\boldsymbol{x}_1), \ldots, \tilde{v}_\varepsilon(\boldsymbol{x}_N))^{\mathrm{T}}$, and $\boldsymbol{\Sigma}_{\tilde{v}_u}$ the covariance matrix of $(\tilde{v}_u(\boldsymbol{x}_1), \ldots, \tilde{v}_u(\boldsymbol{x}_N)^{\mathrm{T}})$. As $\tilde{v}_u(\boldsymbol{x}_i)$ is $v_u(\boldsymbol{x}_i)/(n_i - 1)$ times at $\chi^2(n_i - 1)$ random variable, we have

$$\boldsymbol{\Sigma}_{\tilde{v}_u} = \mathrm{diag}\left(\frac{2v_u^2(\boldsymbol{x}_i)}{n_i - 1}\right) = 2\boldsymbol{V}_u^{[2]}\boldsymbol{E}_2.$$

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

$$\begin{aligned}\boldsymbol{\Sigma}_{\tilde{v}_\varepsilon} &= 2(\boldsymbol{V}_\epsilon + \boldsymbol{V}_u\boldsymbol{E}_1)^{[2]} + \boldsymbol{\Sigma}_{\tilde{v}_u}\boldsymbol{E}_1^{[2]} \\ &= 2\left\{(\boldsymbol{V}_\epsilon + \boldsymbol{V}_u\boldsymbol{E}_1)^{[2]} + \boldsymbol{V}_u^{[2]}\boldsymbol{E}_1^{[2]}\boldsymbol{E}_2\right\}. \qquad (8)\end{aligned}$$

□