AN APPLICATION OF KRIGING WITH NONPARAMETRIC VARIANCE FUNCTION ESTIMATION

J. D. Opsomer, Iowa State University D. Ruppert, Cornell University
M. P. Wand, Harvard University
U. Holst, O. Hössjer, University of Lund
J.D. Opsomer, 222 Snedecor Hall, Ames, IA 50011 (jopsomer@iastate.edu)

Key Words: local linear regression, universal kriging, nitrogen runoff, metamodel.

Abstract:

As part of an on-going agricultural policy research project at Iowa State University, a statistical model was developed for nitrogen runoff from agricultural lands in the Midwest and Northern Plains regions of the U.S. A computation-intensive deterministic model is used to predict runoff for a relatively small set of points and a regression model is fitted to these data. Then, kriging is used to predict at a large number of remaining points in the region of interest. The regression model is comprised of three main components: (1) the mean function, which includes farming practice variables, local soil and climate characteristics and the nitrogen application treatment, is assumed linear in the parameters and fitted by generalized least squares, (2) the variance function, which contains a local as well as 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 simple parametric variogram model to the standardized residuals. The fitting of these three components is iterated until convergence. The model provides an improved fit to the data compared to a previous model that ignored the heteroskedasticity and the spatial correlation.

1. Introduction

Researchers at the Center for Agricultural and Rural Development at Iowa State University (CARD) are developing economic 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. (*see*, among others, Wu *et al.* [10]). In a major departure from previous economic models, the goal of this project is to predict the environmental impacts at both the regional *and* the local level. Local prediction is achieved by using the 128,591 National Resources Inventory (Nusser and Goebel [5]) points and their associated datasets in the region of interest as the basis for the evaluation of pollution impact. 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.* [10]. A map of the study regions containing the locations of the weather stations is given in Figure 1.

Figure 1: Map of the study region (⊕ denotes weatherstation location).

Nitrogen runoff from non-point sources such as

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

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		

Table 1: Model variables.

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 [9]), a widely used (deterministic) biogeophysical process model, provides, at least conceptually, a convenient tool for predicting the nitrogen runoff at the NRI points, both for the current situation as well as for scenarios in which agricultural policies change one or several of the variables in Table 1. Unfortunately, running the model for all NRI points would be prohibitively computer-intensive for establishing a baseline nitrogen runoff level, let alone for evaluating alternative scenarios in which a significant number of points might have changes in their covariate values. It was therefore decided to estimate a statistical "metamodel" on a representative subset of 11,403 datapoint, and use this metamodel in place of EPIC-WQ to predict nitrogen runoff at the remaining observation points, as well as for scenario evaluation. Another advantage of this approach is the estimation of coefficients and accompanying error bands for the covariates, providing additional insights in the nature of the effect of agricultural practices (represented by $NRATE$ and the crop rotation dummy variables in Table 1) on nitrogen pollution.

The original approach of Wu *et al.* [10] was to fit the metamodel by OLS after transforming the dependent variable and adding a limited number of interaction terms. The model was:

$$(YN03)^{1/3.5} = \alpha + Z_1\beta_{z_1} + NRATE * Z_2\beta_{z_2} + X\beta_x + \text{iid errors}, \quad (1)$$

where $X = (LAT, LONG)$ represents the location of the nearest weather station, Z_1 contains the values for the remaining covariates from Table 1 and Z_2 the same except for the removal of the covariate $NRATE$. We will let $Z = [Z_1\ NRATE * Z_2\ X]$ and for simplicity refer to Z as the covariates for this model, and let $\beta = [\beta_{z_1}^T\ \beta_{z_2}^T\ \beta_x^T]^T$. 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 the OLS assumptions in the residuals. Nevertheless, the residuals still exhibited both severe heteroskedasticity, as well as spatial correlation. As noted in Carroll

and Ruppert [1], 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.

The goal of the current paper is to demonstrate how a novel combination of universal kriging and nonparametric variance function estimation can be used to develop an improved regression model for this problem, while maintaining the interpretability of the mean function model (1). The choice of kriging is motivated by the fact that one of primary uses of this model is the prediction of $YN03$ at the large number of points not included in the regression observations, a situation for which kriging has well-known optimality properties (Cressie [2]). Since 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. Since the specific shape of this function was not of particular interest to the CARD researchers, a generalization of the nonparametric variance estimation approach of Ruppert *et al.* [8] is used. This has the advantage of avoiding to introduce bias in the estimation by inappropriate choice of a parametric form for the variance function.

Section 2. proposes a model that explicitly accounts for the heteroskedasticity and spatial correlation in the data, and Section 3. provides a description of the approach used in estimating its various components. In Section 4., the model estimates for the variance and correlation functions are discussed. Section 5. briefly discusses the use of universal kriging for predicting the nitrogen runoff values at the points not included in the metamodel. Details on the fitting procedure as well as the derivations of some of the theoretical results needed for the nonparametric model components can be found in Opsomer *et al.* [6].

2. The Model

The data consist of n_i scalar response measurements Y_{ij} (the $YN03$ measurements from Section 1.) and $q \times 1$ covariates Z_{ij} recorded at $N = 329$ distinct geographic sites x_i, and the total number of observations is denoted by $n = \sum_{i=1}^{N} n_i$.

The model is

$$Y_{ij} = Z_{ij}^T\beta + v_\epsilon(x_i)^{1/2}\varepsilon_i + v_u(x_i)^{1/2}u_{ij} \quad (2)$$

for $j = 1, \ldots, n_i$, $i = 1, \ldots, N$. Here β is a $q \times 1$ vector of parameters, v_ϵ and v_u are bivariate variance functions, the errors u_{ij} are independent and identically distributed with $E(u_{ij}) = 0$, $\mathrm{var}(u_{ij}) = 1$ and the ε_i are such that $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = 1$ and $\mathrm{cov}(\varepsilon_i, \varepsilon_{i'}) = \rho(\|x_i - x_{i'}\|; \theta)$, where $\rho(\cdot; \theta)$ represents a parametric family of stationary, isotropic correlation functions indexed by the (possibly multidimensional) parameter θ. For each i, the u_{ij} are independent of the ε_i.

Since the only available spatial location information is that of the closest weather station, many points share the same "location" x_i, with n_i ranging from 1 to 221. This motivated the inclusion of the first error term in (2). There is also an important 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" statistical packages to perform the estimation and prediction computations for this problem. The remaining errors u_{ij} at a given weather station location x_i were assumed to be independent, since 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*.

3. Estimation Procedure

3.1 Overview

Let $\boldsymbol{Y}$ be the $n \times 1$ vector containing the Y_{ij}'s and $\boldsymbol{Z}$ be the $n \times q$ matrix with (i,j)th row equal to Z_{ij}^T. Let $\boldsymbol{\Sigma}$ represent the variance-covariance matrix of the vector $\boldsymbol{Y}$.

0: (Initialization step) Set $\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}$.

1: Obtain

$$\hat{\beta} = (\boldsymbol{Z}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}.$$

2: Set

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

3: For each x_i such that $n_i \geq 2$ obtain

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

4: Obtain $\hat{v}_u(x_i)$ for all x_i by local linear smoothing of the $\tilde{v}_u(x_i)$.

5: Obtain $\hat{v}_r(x_i)$ by a local linear smoothing of the $\overline{r}_i^2$, and let

$$\hat{v}_\epsilon(x_i) = \hat{v}_r(x_i) - \hat{v}_u(x_i)/n_i.$$

6: Obtain $\hat{\epsilon}_i = \bar{r}_i/\hat{v}_r(x_i)^{1/2}$ and estimate θ in the isotropic correlation model $\rho(\cdot;\theta)$.

7: Obtain

$$\hat{\Sigma} = \hat{\Sigma}_\epsilon + \hat{\Sigma}_u$$

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

$$[\hat{\Sigma}_\epsilon]_{ij,i'j'} = \hat{v}_\epsilon(x_i)^{1/2}\hat{v}_\epsilon(x_i')^{1/2}\rho(\|x_i - x_{i'}\|;\hat{\theta}).$$

This "block" structure for the covariance matrix will allow significant simplifications of the computations in steps 1,4 and 5.

8: Repeat Steps 1-7 R_{iter} times.

3.2 Details on the Implementation

3.2.1 Generalized Least Squares

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

$$\Sigma = \Sigma_u + K^T V_\epsilon K,$$

where Σ_u is a diagonal matrix with repeating "blocks" of length n_i:

$$\Sigma_u = \text{diag}\{v_u(x_i),\ j = 1,\ldots,n_i,\ i = 1,\ldots,N\},$$

V_ϵ is the $N \times N$ covariance matrix of the ϵ_i and K is an $N \times n$ matrix with (i,i') entry equal to 1 for $i' = 1 + \sum_{k=1}^{i-1} n_k, \ldots, \sum_{k=1}^{i} n_k$ and zero otherwise. The inverse of Σ is therefore equal to

$$\Sigma^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}K^T(V_\epsilon^{-1} + K\Sigma_u^{-1}K^T)^{-1}K\Sigma_u^{-1},$$

(Horn and Johnson [4]), which can be rapidly computed since the largest non-diagonal matrix to invert is only $N \times N$.

3.2.2 Variance Function Estimation by Local Polynomial Regression

If we assume normality of the errors, the $\bar{v}_u(x_i)$ in Step 4 are independently distributed, heteroskedastic random variables, with variance equal to $2\bar{v}_u(x_i)^2/(n_i - 1)$, so that the theory developed by Ruppert *et al.* [7] is directly applicable here. The special structure between the estimator and its variance is used in the bandwidth selection of the EBBS algorithm (Ruppert [7]). Details are in Opsomer *et al.* [6].

If we ignore the error caused by the estimation of the mean function, the expectation of $\bar{r}_i^2$ is equal to

$$\text{E}(\bar{r}_i^2) = \text{Var}_\epsilon(x_i) + \frac{\text{Var}_u(x_i)}{n_i},$$

so that the function $v_\epsilon(x_i)$ can be estimated by smoothing the $\bar{r}_i^2$ and subtracting $\text{Var}_u(x_i)/n_i$. The $\bar{r}_i^2$ are correlated random variables, and the estimation of the optimal bandwidth for computing $\hat{v}_\epsilon$ should take that effect into consideration.

Let κ be the $N \times n$ matrix with (i,j) entry equal to $1/n_i$ for $j = 1 + \sum_{k=1}^{i-1} n_k, \ldots, \sum_{k=1}^{i} n_k$ and zero otherwise. Also, let $A \odot B$ denote the elementwise product of equi-sized matrices A and B.

Result 1: Assuming normality of the Y_{ij}'s and ignoring the error due to estimation of the mean $Z\beta$, the covariance matrix of the random vector containing

$$\bar{r}_i^2, \quad i = 1,\ldots,N$$

is given by

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

where $A^{[2]} = A \odot A$.

Because of the special structure of Σ, the elements of this matrix can be seen to be

$$(\kappa\Sigma\kappa^T)_{ii'} = \begin{cases} \frac{v_u(x_i)}{n_i} + v_\epsilon(x_i) & i = i' \\ v_\epsilon(x_i)^{1/2}v_\epsilon(x_{i'})^{1/2}\rho(\|x_i - x_{i'}\|;\theta) & i \neq i' \end{cases}$$

Let $V_u = \text{diag}\{v_u(x_i), i = 1,\ldots,N\}$ and $E = \text{diag}\{1/n_i, i = 1,\ldots,N\}$. The variance-covariance matrix of the $\bar{r}_i^2$ can be written down as

$$2(V_\epsilon + V_u E)^{[2]}.$$

This variance-covariance matrix is used in a bandwidth selection procedure very similar to EBBS and is described in Opsomer *et al.* [6].

3.2.3 Estimation of the Correlation Function by Variogram Fitting

In Step 6, 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 $\bar{r}_i$ are replaced by $\bar{\epsilon}_i = \bar{r}_i/\hat{v}_r(x_i)^{1/2}$. The following model is used:

$$\rho(t;\theta_1,\theta_2) = \left\{1 - \left(\frac{t^2}{1 + t^2/\theta_1} + \theta_2 t\right)\right\}_+$$

with $\theta_1, \theta_2 > 0$. This is the *rational quadratic model* described in Cressie [2] (p.61) with a linear trend added to it and forced to remain postive. The trend was added to improve the fit to the data. Clearly,

other parametric models could be selected as correlation functions for other datasets. The parameters θ_1, θ_2 are estimated by weighted least squares minimization following Cressie [2] (p.96).

The estimate of the spatial variance covariance matrix V_ϵ is computed by setting

$$[\hat{V}_\epsilon]_{ii'} = \rho(x_i - x_{i'}; \hat{\theta})\sqrt{\hat{v}_\epsilon(x_i)\hat{v}_\epsilon(x_{i'})}.$$

4. Results

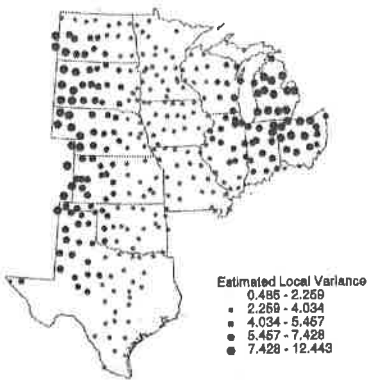

Figure 2: Estimate of the variance function $v_u(\cdot)$ at the weatherstation locations.

The model was run on the CARD dataset, using both the transformed and untransformed nitrogen runoff values as calculated by EPIC-WQ as dependent variables. The transformation was no longer necessary to reduce the heteroskedasticity. This is not too surprising, since the heteroskedasticity was now explicitly accounted for in the model itself. A methodological advantage of the untransformed model is that the predictions computed as in Section 5. are unbiased, while the ones found by using transformed runoff observations are biased after inverting the transformation. We will therefore only show the results for the untransformed model.

Figures 2 and 3 show the nonparametric estimates of the variance functions $v_u(\cdot)$ and $v_\epsilon(\cdot)$ at the weather station locations. Both estimates show a similar pattern of low values in the center and higher ones at the Eastern and Western boundaries of the study region. Most of the variability in the data is explained by the local variance v_u, with the mean

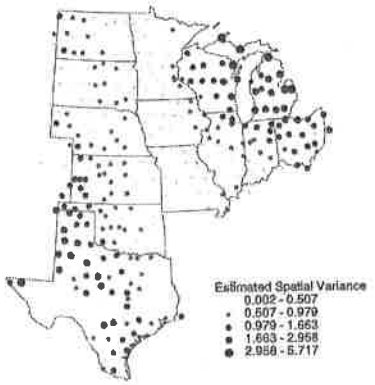

Figure 3: Estimate of the variance function $v_\epsilon(\cdot)$ at the weatherstation locations.

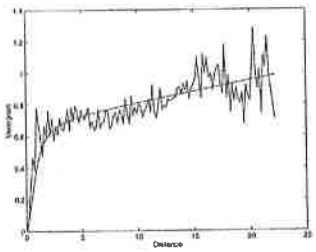

Figure 4: Variogram and estimated variogram function $\hat{\rho}(\cdot; \hat{\theta})$.

value of $\hat{v}_u(x_i)$ equal to 4.418, while that for $\hat{v}_\epsilon(x_i)$ is 0.979. In Figure 4, the variogram of the standardized residuals $\hat{\epsilon}_i$ 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 points at short distances of each other. The estimated variance-covariance matrix $\hat{V}_\epsilon$ was indeed positive definite, with all eigenvalues in the range $[0.005, 81.7]$.

5. Model Predictions

As mentioned in Section 1., the purpose for developing this metamodel is to be able to predict the potential nitrogen runoff at the set of 128,591 NRI

points. If we let Z^* and x^* represent the matrices of covariates and locations at the n^* prediction points, the universal kriging prediction equation for the points $\hat{Y}^* = (\hat{Y}_1^*, \ldots, \hat{Y}_{n^*}^*)^T$ is given by

$$\hat{Y}^* = Z^*\hat{\beta} + C_\epsilon \Sigma^{-1}(Y - Z\hat{\beta}),$$

where C_ϵ is an $n^* \times n$ matrix with elements $[C_\epsilon]_{ii'} = \rho(x_i^* - x_{i'}; \theta)\sqrt{\hat{\nu}_\epsilon(x_i^*)\hat{\nu}_\epsilon(x_{i'})}$ (Cressie [2], p.173). Since the true values for the variance and covariance functions are unknown, we replace them by their estimators obtained by the procedure described in Section 3..

An alternative approach for prediction uses the fact that the prediction and estimation data are at the same set of weather station locations, so that the spatial residuals ε_i can be considered a *lattice* process (Cressie [2]). The vector of spatial errors $e = (\varepsilon_1, \ldots, \varepsilon_N)^T$ can therefore be predicted by a "shrunk" version of the spatial residuals $\bar{r}_i$:

$$\hat{e} = \hat{V}_\epsilon \left(\hat{V}_\epsilon + \hat{V}_{u_n}\right)^{-1} \bar{r},$$

with $\hat{V}_{u_n} = \mathrm{diag}\{\hat{\nu}_u(x_i)/n_i, i = 1, \ldots, N\}$, $\bar{r} = (\bar{r}_1, \ldots, \bar{r}_N)^T$, by a straightforward application of conditional expectations (Fuller [3]). Hence the spatial "correction" $[\hat{C}\hat{\Sigma}^{-1}(Y - Z\hat{\beta})]_i$ at a given location x_i^* can be predicted directly by the corresponding element of the vector $\hat{e}$. This approach is computationally much more efficient than the "full" universal kriging approach described above. Figure 5 shows a plot of the values of the spatial corrections $\hat{\varepsilon}_i$.

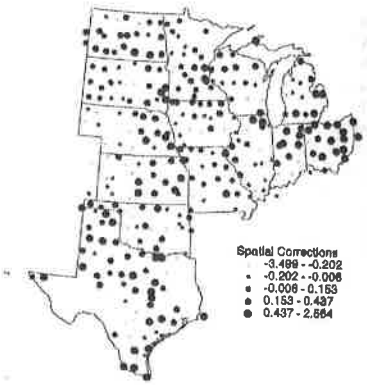

Figure 5: Spatial corrections $\hat{\varepsilon}_i$ at the weatherstation locations.

References

[1] R.J. Carroll and D. Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, New York, 1988.

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

[3] W. A. Fuller. *Measurement Error Models*. John Wiley & Sons, New York, NY, 1987.

[4] R. A. Horn and C. A. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, U.K., 1985.

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

[6] J.-D. Opsomer, D. Ruppert, M.P. Wand, U. Holst, and O. Hössjer. Kriging with nonparametric variance function estimation. Working paper, Center for Agricultural and Rural Development, Iowa State University, 1997.

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

[8] D. Ruppert, M.P. Wand, U. Holst, and Ola Hössjer. Local polynomial variance function estimation. *Technometrics*, 39:262–273, 1997.

[9] A.N. Sharpley and eds. J.R. Williams. Epic-erosion/productivity impact calculator: 1. model documentation. Tech. Bull. 1768, USDA, Washington, DC, 1990.

[10] J. Wu, P.G. Lakshminarayan, and B.A. Babcock. 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 1996.