Multivariate Geostatistics for Precision Agriculture

Noel Cressie
The Ohio State University, Department of Statistics
Columbus, OH 43210-1247 U.S.A.
ncressie@stat.ohio-state.edu

Jay M. Ver Hoef
Alaska Department of Fish and Game
1300 College Road
Fairbanks, AK 99701 U.S.A.
ffjmv@uaf.edu

1. Introduction

In very recent times, agricultural practices have begun to exploit within-field heterogeneity to obtain higher yielding crops and more environmentally friendly fertilizer- and pesticide-application schemes. Global positioning systems installed in new farm equipment can measure the application of fertilizer and insecticides down to meters, and they provide closely spaced data on crop yields. The possibility of characterizing within-field heterogeneity of a multitude of variables has brought *precision agriculture* (PA) to the modern American farm. One of the goals of PA is to map variables, such as soil properties, so that the farmer can apply fertilizer or insecticides sparingly. Because yield data is plentiful and typically correlates with other variables that are difficult to measure, it is reasonable to use information on past yield to make better maps of soil and entomological variables, for example. The area of study known as geostatistics not only provides maps of predictions from noisy, incomplete data, but also maps of prediction standard errors. In this paper, we apply a multivariate geostatistical technique known as cokriging and compare it to the univariate geostatistical technique known as kriging, for some within-field measurements on corn yield and soil pH in Iowa, U.S.A.

2. Kriging and Cokriging

Spatial statistics and, in particular, geostatistics is concerned with sampling from a spatial domain. Let the kth spatial random variable at location $\mathbf{s}$ be denoted $Z_k(\mathbf{s})$. The cross-variogram (Cressie, 1991, pg. 140) is defined as,

$$2\gamma_{km}(\mathbf{h}) \equiv var(Z_k(\mathbf{s}) - Z_m(\mathbf{s}+\mathbf{h})); \tag{1}$$

when $k = m$, we call it the variogram. Suppose there are only two variable types and let the kth variable type consist of n_k observed locations given by the random vector $\mathbf{Z}_k \equiv (Z_k(\mathbf{s}_{k,1}), Z_k(\mathbf{s}_{k,2}), ..., Z_k(\mathbf{s}_{k,n_k}))'$; $k = 1, 2$. Define the data vector $\mathbf{Z} \equiv (\mathbf{Z}_1', \mathbf{Z}_2')'$, and suppose we want to predict $Z_1(\mathbf{s}_0)$. Assume that $E(Z_k(\mathbf{s})) = \mathbf{x}_k(\mathbf{s})'\boldsymbol{\beta}_k$, where $\mathbf{x}_k(\mathbf{s})$ are known explanatory variables; $k = 1, 2$. Cokriging prediction equations for $Z_1(\mathbf{s}_0)$ are given as $p_1(\mathbf{s}_0; \mathbf{Z}) = \boldsymbol{\lambda}'\mathbf{Z}$, where the formula for $\boldsymbol{\lambda}$ in terms of (1) is given in Ver Hoef and Cressie (1993). Associated with the cokriging predictor $\mathbf{p}_1(\mathbf{s}_0; \mathbf{Z})$ is the cokriging variance $\sigma_1^2(\mathbf{s}_0)$, whose formula in terms of (1) is again found in Ver Hoef and Cressie (1993).

In all that is to follow, we assume that $E(Z_k(\mathbf{s})) \equiv \mu_k$, and hence $\mathbf{x}_k(\mathbf{s}) \equiv 1$; $k = 1, 2$. However, our analyses are easily modified to the more general situation of nonconstant mean. Next, we need to choose variogram and cross-variogram models. Ver Hoef and Barry (1996), Ver Hoef and Barry (1998), and Ver Hoef, Cressie, and Barry (2001) show how to use the Fast Fourier Transform (FFT) to fit any moving-average function

that can be approximated by many small rectangles. That is,

$$\gamma_{jj}(\mathbf{h}|\theta_j,\nu_j) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_j^2(\mathbf{u}|\theta_j) - g_j(\mathbf{u}|\theta_j)g_j(\mathbf{u}-\mathbf{h}|\theta_j)d\mathbf{u} + \nu_j^2\,, \tag{2}$$

for $\mathbf{h} \neq \mathbf{0}$, and zero otherwise; $\gamma_{jj}(\mathbf{h}|\theta_j,\nu_j)$ is the semivariogram and $g_j(\mathbf{u}|\theta_j)$ is called the moving-average function. And the semicross-variogram is

$$\begin{aligned}\gamma_{jk}(\mathbf{h}|\boldsymbol{\theta},\boldsymbol{\nu},\boldsymbol{\rho},\boldsymbol{\Delta}) \;=\; & (1/2)\Bigg\{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_j^2(\mathbf{u}|\theta_j)d\mathbf{u} + \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_k^2(\mathbf{u}|\theta_k)d\mathbf{u} \\ & -2\rho_j\rho_k\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_j(\mathbf{u}|\theta_j)g_k(\mathbf{u}-\mathbf{h}+\boldsymbol{\Delta}_k-\boldsymbol{\Delta}_j|\theta_k)d\mathbf{u} + \nu_j^2+\nu_k^2\Bigg\}.\end{aligned} \tag{3}$$

The moving-average constructions (2) and (3) are very flexible because the moving-average functions $g_j(\mathbf{u}|\theta_j)$ and $g_k(\mathbf{u}|\theta_k)$ have no restrictions other than square integrability. Because the semivariograms (2) and semi-crossvariograms (3) are built from a valid multivariate spatial process, they yield valid kriging and cokriging equations.

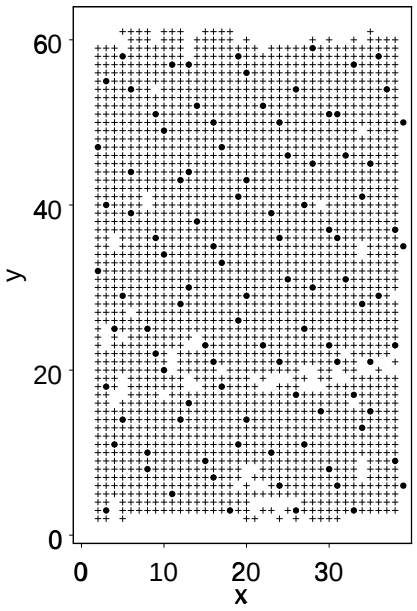

Figure 1. Spatial locations of data. Yield locations are shown with crosses and pH locations are shown with open circles

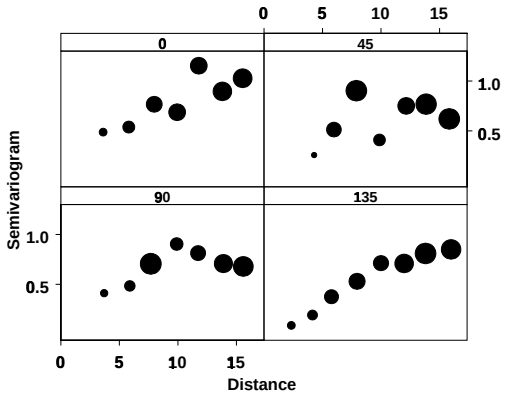

Figure 2. Empirical semivariograms for pH. Panel numbers indicate semivariograms, with directions of 0, 45, 90, and 135 degrees from north. The larger the circle, the larger the number of pairs of points for that bin in the empirical semivariogram.

3. Applying Cokriging to Precision-Agriculture Data

We analyze precision-agriculture data from McGarvey Field in central Iowa, using some new models developed for cokriging. These data were first discussed by Chen and Breidt (1997), who carried out a different spatial analysis using nongeostatistical models; however, no field maps of pH were given. The spatial locations of the data are shown in Fig. 1; there are 2068 yield values and 100 pH values. Prior to analysis, we transformed the yield values to approximate normality by taking the 6th power; we then standardized the transformed data

by subtracting off their mean and dividing by their standard deviation. The empirical semivariogram (Cressie, 1993, p. 69) for the pH data is shown in Fig. 2. There appears to be spatial dependence in pH, and it seems to be strongest at the 135-degree angle. Fig. 3 shows strong spatial dependence for yield as well, but it appears to be equal in all directions (isotropy). The empirical semicross-variogram shows larger values for small lags (Fig. 4), which indicates negative cross-correlation. Also notice that we show Fig. 4 in eight directions rather than four because, although the variogram is radially symmetric, the cross-variogram is not. Moving-average models were fit to the data using the methods described by Ver Hoef, Cressie, and Barry (2001). Using the fitted models, the semivariogram and semicross-variogram values were used to predict pH values for a grid of points ranging from $1 \leq x \leq 39$ and $1 \leq y \leq 61$, using both kriging and cokriging. A map of the predicted values using cokriging is given in Fig. 5; the kriging results looked quite similar. A map of cokriging standard errors is given in Fig. 6.

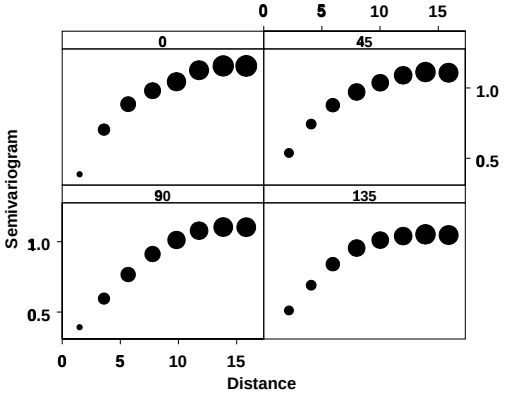

Figure 3. Empirical semivariograms for yield. Panel numbers indicate directional semivariograms, with directions of 0, 45, 90, and 135 degrees from north. The larger the circle, the larger the number of pairs of points for that bin in the empirical semivariogram.

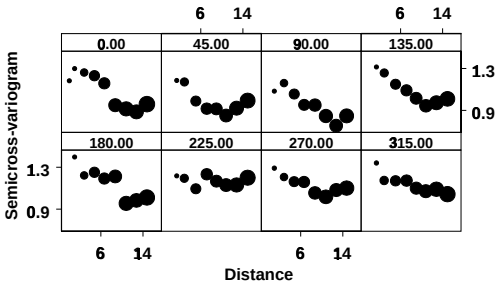

Figure 4. Empirical semicross-variograms for yield. Panel numbers indicate directional semicross-variograms, with directions of 0, 45, 90, 135, 180, 225, 270, and 315 degrees from north. See Figure 3 caption for an explanation of circle size.

To check the model, we used some summary graphs and statistics based on *cross-validation*. Let $\widehat{Z}_1(\mathbf{s}_{1i})$ be the (co)kriged value of the pH variable at location $\mathbf{s}_{1i}$ when the ith datum for pH at location $\mathbf{s}_{1i}$ has been *removed* from the data set; we momentarily suppress the difference in notation between cokriging and kriging. Fig. 7 shows all predicted and actual values for both kriging and cokriging. Notice that both kriging and cokriging produce predictions that match well with the observed values. We checked for bias in both kriging and cokriging by calculating $(1/n_1)\sum_{i=1}^{n_1}(\widehat{Z}_1(\mathbf{s}_{1i}) - Z_1(\mathbf{s}_{1i}))$. We used $RMSPE \equiv \left\{\sum_{i=1}^{n_1}(\widehat{Z}_1(\mathbf{s}_{1i}) - Z_1(\mathbf{s}_{1i}))^2/n_1\right\}^{1/2}$, the cross-validation root-mean-squared-prediction error, to assess the prediction performance of both kriging and cokriging. Let $\widehat{var}(\widehat{Z}_1(\mathbf{s}_{1i}))$ denote the estimated (co)kriging variance at location $\mathbf{s}_{1i}$ when the ith datum for pH at location $\mathbf{s}_{1i}$ has been removed from the data set. Then define $RMEV \equiv \{\sum_{i=1}^{n_1}(var(\widehat{Z}_1(\mathbf{s}_{1i}))/n_1\}^{1/2}$. If the estimated prediction variances are correct, then $RMEV$ should be close to $RMSPE$. We also wanted to assess whether the estimated prediction standard errors were valid. If we denote $\widehat{se}(\widehat{Z}_1(\mathbf{s}_{1i})) \equiv (\widehat{var}[\widehat{Z}_1(\mathbf{s}_{1i})])^{1/2}$, then the prediction interval, $80\%PI \equiv \sum_{i=1}^{n_1} \mathcal{I}[|\widehat{Z}_1(\mathbf{s}_{1i}) - Z_1(\mathbf{s}_{1i})| < 1.28 \cdot \widehat{se}(\widehat{Z}_1(\mathbf{s}_{1i}))]$, should be about 0.80. For kriging, we used the fitted parameters for the pH model only, and for cokriging we used all fitted parameters. The results for cokriging (kriging) are as follows. Bias: 0.032(0.035); RMSPE: 0.586(0.592); RMEV: 0.579(0.591); and 80%PI: 0.820(0.800). We conclude from these results that there is negligible bias; based on RMSPE and RMEV, there is a 1 to 2% improvement when using cokriging over kriging, and both methods appear to have statistically valid prediction variances and prediction intervals.

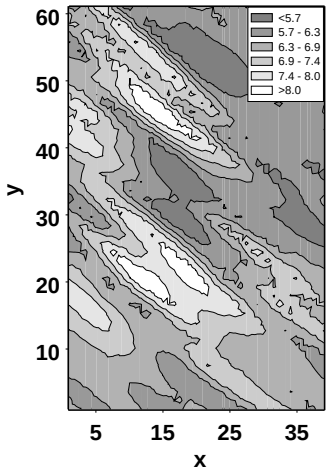

Figure 5. Cokriging predictions of soil pH using yield as a secondary variable.

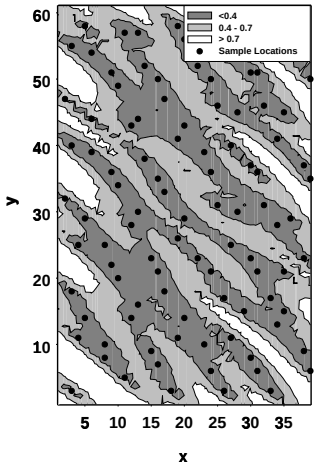

Figure 6. Cokriging prediction standard errors of soil pH using yield as a secondary variable.

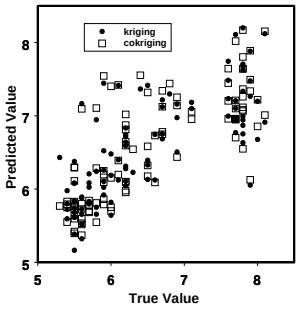

Figure 7. Cross-validation plots for cokriging (squares) and kriging (dots).

Acknowledgment: Support was provided by the U.S. EPA under grant no. R827-257-01-0.

REFERENCES

[1] Barry, R.P. and Ver Hoef, J.M. 1996. Blackbox kriging: spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics* **3**: 1-25.

[2] Chen, C. and Breidt, F.J. 1997. A spatial autoregressive model in precision farming. *ASA Proceedings of the Section on Statistics and the Environment, American Statistical Association, Alexandria, VA*, 37-42.

[3] Cressie, N. 1993. *Statistics for Spatial Data, Revised Edition.* John Wiley and Sons, New York, 900 pp.

[4] Ver Hoef, J.M. and Cressie, N. 1993. Multivariable spatial prediction. *Mathematical Geology* **25**: 219-240. [Erratum, *Mathematical Geology* **26**: 273-275 (1994).]

[5] Ver Hoef, J.M. and Barry, R.P. 1998. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference* **69**: 275 - 294.

[6] Ver Hoef, J.M., Cressie, N., and Barry, R.P. 2001. Flexible spatial models for kriging and cokriging using moving averages and the Fast Fourier Transform (FFT). Technical Report No. 675, Department of Statistics, The Ohio State University.