where F is $n \times n$ and orthogonal, that is, $F'F = FF' = I_n$; $F_1'F_1 = I_t$ and $F_1'F_2 = 0$; $F_2'F1 = 0$, G is $n \times t$ and zero below the main diagonal, F_1 is $n \times t$ and F_2 is $n \times (n - t)$, and G_1 is an upper triangular matrix of size $t \times t$. Then compute the following Choleski decomposition:

$$F_2'K_sF_2 = LL' \tag{A2}$$

where L is a lower $t \times t$ triangular matrix.

The singular value decomposition (SVD) of L is:

$$L = \mathrm{UDV}' \tag{A3}$$

The influence matrix "A" as a function of c_0 can then be written as:

$$A(c_0) = F_1F_1' + F_2UD^2(D^2 + c_0)^{-1}U'F_2' \tag{A4}$$

Lengthy but otherwise not difficult calculations show that GCV can be written as:

$$\mathrm{GCV}(c_0) = \frac{n \sum_{i=1}^{n-t} \left| \frac{c_0}{d_i^2 + c_0} \right| y_i^2}{\left[\sum_{i=1}^{n-t} \frac{c_0}{d_i^2 + c_0} \right]^2} \tag{A5}$$

where

$$y = U'F_2'z \tag{A6}$$

Mathematical Geology, Vol. 27, No. 5, 1995

Inner Product Matrices, Kriging, and Nonparametric Estimation of Variogram[1]

Subhash Lele[2]

Two important problems in the practical implementation of kriging are: (1) estimation of the variogram, and (2) estimation of the prediction error. In this paper, a nonparametric estimator of the variogram to circumvent the problem of the precise choice of a variogram model is proposed. Using orthogonal decomposition of the kriging predictor and the prediction error, a method for selecting, what may be considered, a "statistical neighborhood" is suggested. The prediction error estimates based on this scheme, in fact, reflects the true prediction error, thus leading to proper coverage for the corresponding prediction interval. By simulations and a reanalysis of published data, it is shown that the proposals made in this paper are useful in practice.

KEY WORDS: conditionally negative definite function, positive definite matrix, prediction error principal components.

INTRODUCTION

Various scientific disciplines require the collection and prediction of data over space. For example, in mining where the goal is to predict ore concentrations over the entire study area, samples are collected at various locations. To predict concentration at locations where the samples are not collected, geostatistics uses a technique known as kriging. Kriging produces a map of ore concentrations for the entire site which can be used for planning and operating mining activities. This same technique has applications in environmental data collection where the goal is to predict environmental degradation or clean-up based on the data collected at a discrete number of monitoring locations at a site. As in mining, a useful tool for site assessment and clean-up of a contaminated site is a contour map of contaminant concentrations over the area of interest. Environmental decision makers then could use this map to identify those areas which should be excavated to protect public health, those which pose little or no risk, and those where the uncertainty is large enough to warrant additional sampling.

[1]Received 25 June 1993; revised 14 December 1994.

[2]Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Maryland 21205.

0882-8121/95/0700-0673$07.50/1

The attraction of the kriging procedure in these applications is twofold. First, it offers a statistical justification for the way it takes point data (data from locations that have been sampled) and generates a smooth, interpolated map (e.g., a contour plot) of contaminant concentrations. This is in contrast to the other conventional methods of two-dimensional smoothing and interpolation that generally are acknowledged to be *ad hoc*, although effective. Second, the kriging procedure generates explicit uncertainty measures (e.g., prediction intervals) for the interpolated and smoothed estimates—both for estimates of concentrations at particular locations, and for estimates of averages within a defined area. These uncertainty measures can be used for building precise margins of safety into a decision rule that uses the estimate (e.g., a decision rule for guiding a clean-up for the boundaries of a contaminant source) and they can be used in more sophisticated adaptive procedures (e.g., a value-of-information approach) for making decisions about collecting additional data and for optimal selection of sampling locations for additional data.

We refer the reader to Cressie (1991) or Journel and Huijbregts (1978) for detailed mathematical as well as applied description of the kriging technique. However, the basic idea behind kriging is easy to understand. Suppose one wants to predict a contaminant concentration at a new, unsampled location. It intuitively makes sense to consider a weighted linear combination of the observed contaminant concentrations as a predictor. It also makes sense to give more weight to those observations at locations that are more similar to the location at which prediction is desired, than the ones which are different. Kriging thus needs (Journel, 1988):

(1) identification of a "dissimilarity" or "distance" measure between locations. The variogram model provides this; and

(2) identification of "optimal" weights based on the variogram model.

Kriging, as typically implemented (assuming that the data are detrended properly), thus requires the practitioner to select a variogram model. Moreover, because the parameters in the selected model usually are unknown, one has to decide the method for the estimation of these parameters (e.g., ordinary least squares, weighted least squares, maximum likelihood, etc.) and then, based on these estimated parameters, predict the unknown concentrations along with the prediction error associated with them. If n, the number of sampled locations, is large, one has to decide the kriging neighborhood (a subset of the total sample) to reduce the computational burden.

The goal of this paper is several fold.

(1) Recently there have been several attempts at eliminating the selection of variogram model step in kriging through the use of nonparametric estimators of variogram (Shapiro and Botha, 1991; Hall, Fisher, and Hoffman, 1994; Cherry, 1994). In this paper, an alternative nonparametric estimator which is computationally easy is suggested.

(2) There have been several suggestions on the use of orthogonal basis for kriging prediction (Journel, 1977; Kacewicz, 1991; Vecchia, 1992). The usual definition of kriging neighborhood is in terms of "geographical nearness". We suggest to define neighborhood in terms of statistical nearness. The use of orthogonal basis for the selection of a "statistical kriging neighborhood" is recommended. The number of terms in the kriging predictor (defined in terms of the orthogonal basis) is determined in a sequential fashion. The computational burden is demonstrated to be substantially smaller than the standard kriging procedure. Moreover it is demonstrated that the predictor based on the first few orthogonal terms is reasonably close to the optimal predictor. Thus, we do not lose statistical efficiency, but stand to gain substantially in computational simplicity.

(3) Kriging involves not just the point prediction of an observation at a new location but also, and perhaps more importantly, the uncertainty (i.e., prediction error) associated with it. Zimmerman and Cressie (1992) discuss rigorously the problem of estimation of prediction error based on the estimated variogram. They show that, in general, estimated mean squared prediction error is too optimistic (smaller than the actual mean squared prediction error) and hence the associated prediction intervals, in general, have smaller than nominal coverage. In this paper, it is demonstrated that if the kriging predictor is based only on the first few orthogonal terms, it tends to estimate the true prediction error accurately and hence obtain prediction intervals which have coverage close to the nominal coverage. Thus, although these predictors are suboptimal, they tend to be reflective of the truth related to the statement about the uncertainty measure. This is a highly important feature for the environmental management decision-making process.

NOTATION AND PRELIMINARY RESULTS

The following notation is used throughout the paper.

(1) $s_1, s_2, \ldots, s_n$ denote the locations of sites where the process is observed.

(2) $z(s_1), z(s_2), \ldots, z(s_n)$ denote the observations at the corresponding locations. $Z(s_1), Z(s_2), \ldots, Z(s_n)$ denote the random variables at those locations. In general, capital letters denote the random variables and the corresponding small letters denote the realization of those random variables.

(3) $Z' = (Z(s_1), Z(s_2), \ldots, Z(s_n))$ denotes the vector of random variables. Vectors are considered column vectors. Vectors with superscript "t" are transposed vectors, and hence are row vectors.

(4) $Z_c' = (Z(s_2) - Z(s_1), Z(s_3) - Z(s_1), \ldots, Z(s_n) - Z(s_1))$ denotes the vector of contrasts.

(5) Ψ denotes the variance–covariance matrix of the vector of contrasts Z_c. This is a square, symmetric matrix of dimension $n - 1$ by $n - 1$. Thus the (i, j)th entry in this matrix corresponds to cov $(Z(s_i) - Z(s_1), Z(s_j) - Z(s_1))$ where $i = 2, 3, \ldots, n$ and $j = 2, 3, \ldots, n$. Covariance between two random variables is denoted by "cov." Ψ is termed the "Inner product variogram matrix."

(6) The variogram value between locations i and j is denoted by $2\gamma_{ij}$. Thus, $2\gamma_{ij}$ = var $(Z(s_i)-Z(s_j))$ where "var" denotes the variance of a random variable. Here i and $j = 1, 2, \ldots, n$.

(7) The "variogram matrix" is denoted by Γ, thus the (i, j)th entry in Γ is $2\gamma_{ij}$.

(8) The new location, at which the prediction is desired, is denoted by s_0. The random variable to be predicted is denoted by $Z(s_0)$.

(9) The vector containing the values cov $(Z(s_i) - Z(s_1), Z(s_0) - Z(s_1))$, $i = 2, 3, \ldots, n$ is denoted by ψ_0. This is of length $(n - 1)$.

(10) The vector containing the values var $(Z(s_i) - Z(s_0))$, $i = 1, 2, 3, \ldots, n$ is denoted by γ_0. This is of length n.

Following are some preliminary results and properties of the Ψ matrix.

Result 1. If the underlying process is intrinsically stationary, Ψ matrix exists.

Proof. From the basic properties of Hilbert space with an inner product $\langle ., . \rangle$, we know that: $2\langle u, v \rangle = \langle u, u \rangle + \langle v, v \rangle - \langle u - v, u - v \rangle$. If we define covariance between two vectors as the inner product, then it follows that:

$$2\text{cov}\,(Z(s_i) - Z(s_1), Z(s_j) - Z(s_1)) = \text{var}\,(Z(s_i) - Z(s_1)) + \text{var}\,(Z(s_j) - Z(s_1)) - \text{var}\,(Z(s_i) - Z(s_j))$$

or equivalently,

$$\Psi_{ij} = \gamma_{i1} + \gamma_{j1} - \gamma_{ij} \tag{1}$$

Because the underlying process is intrinsically stationary (Cressie, 1991 p. 60–61), the right-hand side exists and hence Ψ exists. The following relationship follows immediately from (1)

$$2\gamma_{ij} = \Psi_{ii} + \Psi_{jj} - 2\Psi_{ij} \tag{2}$$

Result 2. The matrix Ψ is necessarily a positive semidefinite matrix.

This follows immediately from the observation that Ψ is a variance covariance matrix.

Result 3. Ψ can be written as $\Psi = PDP'$ where D is a diagonal matrix of eigenvalues $d_2, d_3, \ldots, d_n$ and P is a matrix of eigenvectors of Ψ.

This is just a spectral decomposition of a real, symmetric matrix. Note that because Ψ is positive semidefinite, the eigenvalues d_i are nonnegative, although some of them may be zero.

ORDINARY KRIGING IN TERMS OF INNER PRODUCT VARIOGRAM

Ordinary kriging refers to spatial prediction under the following two assumptions (Cressie, 1991, p. 120).

(1) For all s in the domain D of the process,

$$Z(s) = \mu + \delta(s)$$

where μ is an unknown real number representing the mean of the process and $\delta(\cdot)$ is a zero mean, intrinsically stationary process.

(2) The predictor $p(Z; s_0)$ is such that:

$$p(Z; s_0) = \sum^{n} \lambda_i Z(s_i)$$

where $\Sigma_{i=1}^{n} \lambda_i = 1$.

The second condition guarantees uniform unbiasedness of the predictor. We want to determine a set of λs such that the mean-squared prediction error is minimized. The mean-squared prediction error is given by:

$$\sigma_e^2 = E(Z(s_0) - p(Z; s_0))^2$$

We now derive the solution set λ. Because $\Sigma_i^n \lambda_i = 1$, we can write

$$Z(s_0) = \sum^{n} \lambda_i (s_i \quad (Z(s_0) \quad Z(s_1)) \quad \sum^{n}_{i=2} \lambda_i (Z(s_i) \quad Z(s_1))$$

Here it is important to note that $\lambda_2, \quad ., \lambda_n$ are unconstrained. The predictor is written as: $Z(s_1) + \Sigma_{i=2}^{n} \lambda_i (Z(s_i) \quad Z(s_1))$. It now is easy to see that

$$E(Z(s_0) - \sum_{i=1}^{n} \lambda_i Z(s_i))^2 = E(Z(s_0) - Z(s_1))^2 + \sum_{i=2}^{n} \lambda_i^2 E(Z(s_i) - Z(s_i))^2$$

$$+ \sum_{i \neq j} \lambda_i \lambda_j E((Z(s_i) - Z(s_1))(Z(s_j) - Z(s_1)))$$

$$- 2 \sum_{i=1}^{n} \lambda_i E((Z(s_i) - Z(s_1))(Z(s_0) - (s_1)))$$

$$= 2\gamma_{10} + \sum_{i=2}^{n} \lambda_i^2 \Psi_{ii} + \sum_{i \neq j} \lambda_i \lambda_j \Psi_{ij} - 2 \sum_{i=2}^{n} \lambda_i \psi_{i0}$$

Differentiating with respect to λ_i and equating the derivatives to zero, we get

$$\lambda_i \Psi_{ii} + \sum_{j \neq i} \lambda_j \Psi_{ij} = \psi_{i0}$$

for $i = 2, 3$, n. In matrix notation, this can be written as

$$\Psi \lambda = \psi_0$$

where $\lambda' = (\lambda_2, \lambda_3, \ldots, \lambda_n)$.

Provided that Ψ is positive definite, the optimal set of λs is given by

$$\lambda = \Psi^{-1} \psi_0$$

If Ψ is only positive semidefinite, it implies that some of the contrasts in Z_c are linearly related to each other. This implies that some of the λs can be equated to zero without any loss.

This last comment points to the possibility of using the principal components (Jolliffe, 1986) of Z_c as predictors instead of the individual components separately. We explore this possibility next. Let P be the matrix of eigenvectors of Ψ with $P_{(i)}$, the ith eigenvector. Then

$$P'_{(i)} Z_c = \sum_{j=2}^{n} P_{ij}(Z(s_j) - Z(s_1))$$

is the ith principal component of Z_c. For notational simplicity it will be denoted by y_i. We consider the predictor:

$$P(Z; s_0) = Z(s_1) + \sum_{i=2}^{n} \lambda_i^* y_i$$

Using the results about the principal components, namely $E(y_i) = 0$ and $E(y_i, y_j) = 0$ if $i \neq j$ and $E(y_i^2) = d_i$, one can write:

$$E\left((Z(s_0) - Z(s_1) - \sum_{i=2}^{n} \lambda_i^* y_i\right)^2 = E(Z(s_0) - Z(s_1))^2 + \sum_{i=2}^{n} (\lambda_i^*)^2 d_i$$

$$- 2 \sum_{i=2}^{n} \lambda_i^* E(y_i(Z(s_0) - Z(s_1)))$$

$$= 2\gamma_{10} + \sum_{i=2}^{n} (\lambda_i^*)^2 d_i - 2 \sum_{i=2}^{n} \lambda_i^* P'_{(i)} \psi_0 \tag{3}$$

Differentiating with respect to λ_i^*s, equating to zero and writing in the matrix notation, we get:

$$D\lambda^* = P' \psi_0$$

The optimal λ^*s are thus given by:

$$\lambda_i^* = P'_{(i)} \psi_0 / d_i$$

$d_i > 0$. If $d_i = 0$, corresponding λ_i^*s are defined to be zero.

Prediction Error for p(Z; s_0). Using Equation (3), it is easy to write the prediction error for $p(Z; s_0)$ as:

$$\sigma_e^2 = 2\gamma_{10} + \sum_{i=2}^{n} (\lambda_i^*)^2 d_i - 2 \sum_{i=2}^{n} \lambda_i^* P'_{(i)} \psi_0$$

$$= 2\gamma_{10} - \sum_{i=2}^{n} (P'_{(i)} \psi_0)^2 / d_i \tag{4}$$

Let us look at the interpretation of the decomposition of the prediction error given in Equation (4).

(1) Suppose only one of the observations $\{Z(s_1), Z(s_2), \ldots, Z(s_n)\}$ is used to predict $Z(s_0)$. Then, it is obvious that if $Z(s_1)$ is used as the predictor of $Z(s_0)$, the kriging prediction error would be $2\gamma_{10}$.

(2) $(P'_{(i)} \psi_0)^2 / d_i$s indicate the successive reduction in the prediction error because of the inclusion of the ith principal component in the predictor function.

(3) The addition of the principal components can be done conveniently in a *sequential* fashion. Note that each λ_i^* depends only on the ith eigenvector of the matrix Ψ. For large matrices, it is a relatively simple numerical analysis problem to evaluate eigenvalues and eigenvectors in a sequential fashion (Golub and van Loan, 1989).

A scheme for selecting the number of principal components to be included in the predictor $p(Z; s_0)$, using the decomposition in Equation (4) and the comments, now can be prescribed.

Given Ψ, ψ_0 and γ_{10}, calculate:

Predictor	Prediction error	
$p(Z; s_0) = Z(s_1)$	$\sigma_e^2 = 2\gamma_{10}$	(5)
$p(Z; s_0) = Z(s_1) + \lambda_2^* y_2$	$\sigma_e^2 = 2\gamma_{10} - (P'_{(2)}\psi_0)^2/d_i$	

Stop adding principal components to the predictor when the prediction error σ_e^2 is smaller than a prescribed number, or when the incremental improvement in σ_e^2 is smaller than a prescribed number or when d_is become negligible. This is what we consider "statistical neighborhood." See Davis and Grivet (1984) for an interesting discussion of local versus global kriging neighborhood. If s_1 is fixed, the neighborhood discussed here corresponds to the "global neighborhood." Note that we are indexing the eigenvalues and eigenvectors of Ψ from 2 to n and NOT 1 to $(n - 1)$.

NONPARAMETRIC ESTIMATION OF VARIOGRAMS

In the development of the previous section it is assumed that the variogram and hence the inner product variogram matrices are known. Of course, in practice, they are seldom known and need to be estimated using the available data. Various estimators, both parametric and nonparametric, have been suggested in the literature. See Cressie (1991, chapter 2) or Zimmerman and Zimmerman (1991) for a survey of the available methodologies and theoretical results associated with them.

Most of the estimation procedures are based on the assumption that only one realization from the underlying process is available. Therefore, we have to assume certain stationarity and ergodicity properties of the underlying process and use the spatial replication to estimate the variogram. Another assumption is that of isotropy implying that the variogram depends only on the distance between two locations and not on the direction. Geometric anisotropy may be included in the parametric model. Though the development of the previous section is independent of the assumptions of stationarity and isotropy, for any data analytic applications we need to assume these properties. For the purpose of this paper, we will assume stationarity and isotropy.

If the underlying process is stationary and isotropic, several different estimators of variogram are available (see Cressie, 1991, chapter 3; Journel, 1988).

For the results described here, all that is needed is a consistent estimator of the variogram. In the following, the classical or moment estimator of the variogram is described.

Moment Estimator of the Variogram. Given the observations $z(s_1), z(s_2), \ldots, z(s_n)$, the moment estimator of the variogram at a distance h is given by:

$$\widehat{2\gamma(h)} = |N(h)|^{-1} \sum_{N(h)} (Z(s_i) - Z(s_j))^2$$

where

$$N(h) = \{(s_i, s_j): \|s_i - s_j\| = h; \quad i, j = 1, 2, \ldots, n\}$$

and $|N(h)|$ is the number of distinct pairs in $N(h)$, $\|s_i - s_j\|$ is the Euclidean distance between the locations s_i and s_j. If the data are spaced irregularly, the variogram estimator usually is smoothed according to some tolerance region. One also can simply smooth the scatter plot of $(z(s_i) - z(s_j))^2$ vs. $\|s_i - s_j\|$ using any kernel smoothing or spline smoothing method (Silverman, 1986).

Unfortunately, such a smoothed variogram is not guaranteed to be conditionally negative definite and hence in general, would not be a valid variogram (Chritakos, 1984). Attempts have been made for fitting a function from the class of conditionally negative definite functions (Cherry, 1994; Shapiro and Botha, 1991; Hall, Fisher, and Hoffman, 1994) to this scatter. These definitely are of great interest both to those who use kriging for prediction and to those who use variograms to characterize spatial dependencies. However, the class of conditionally negative definite functions (Schoenberg, 1938; Chritakos, 1984) is difficult mathematically to handle, especially for two and higher dimensional data. For example, the basis functions in a two-dimensional situation are Bessel functions. These, themselves, are infinite series. Thus approximating an estimated variogram (which need not be conditionally negative definite) by a valid conditionally negative definite function is a difficult mathematical and computational task.

The class of positive definite functions (being a dual) is equally difficult to handle. However, given a finite dimensional matrix, it is comparatively easy to approximate it with a positive definite matrix (Rousseeuw and Molenberghs, 1993; Devlin, Gnanadesikan, and Kettenring, 1975; Mead, 1991; Goulard and Volz, 1992). We propose to transform the problem of estimation of the valid variogram matrix (which is conditionally negative definite) to the problem of estimation of the valid inner product variogram matrix (which needs to be a positive definite matrix).

In the following we describe steps to obtain a nonparametric estimator of variogram. An argument for the validity, that is the conditional negative definiteness, of the resultant estimator will follow.

Step 1. Calculate the squared differences $\frac{1}{2}(z(s_i) - z(s_j))^2$. Calculate the

moment estimator of the variogram with the constraint that each bin contains at least 30 observations. This is an accepted constraint (e.g., Journel and Huijbregt, 1978). Plot the moment estimator against the distances.

Step 2. Use a spline function to smooth the given plot. The generalized cross-validation criterion is used to select the smoothing parameter (Silverman, 1984; Sahba, 1990).

Step 3. Based on this smoothed function, calculate $\tilde{\gamma}_{ij}$s for distances $\|s_i - s_j\|$s. The corresponding inner product variogram values are obtained by utilizing the relationship

$$\tilde{\Psi}_{ij} = \tilde{\gamma}_{i1} + \tilde{\gamma}_{j1} - \tilde{\gamma}_{ij}$$

Step 4. Construct the matrix $\tilde{\Psi}$ based on $\tilde{\Psi}_{ij}$s. Calculate the positive definite approximation to the matrix $\tilde{\Psi}$.

The following is a widely used procedure to obtain a positive definite approximation to a given matrix, such that the sum of squared differences between the elements of the two matrices is minimized (Mead, 1992a; for other solutions, see Rousseeuw and Molenberghs, 1993; Devlin, Gnanadesikan, and Kettenring, 1975.)

(a) Calculate the eigenvalues and eigenvectors of $\tilde{\Psi}$. Write the spectral decomposition of $\tilde{\Psi} = P\tilde{D}P^T$.

(b) Let D^* be a diagonal matrix of the same dimension as $\tilde{D}$, such that all the eigenvalues smaller than a prespecified value $\epsilon > 0$ are replaced by ϵ.

(c) Let $\hat{\Psi} = PD^*P^T$.

Then $\hat{\Psi}$ is necessarily a positive definite matrix.

Step 5. Obtain $\hat{\gamma}_{ij}$s from $\hat{\Psi}_{ij}$ using the relationship:

$$2\hat{\gamma}_{ij} = \hat{\Psi}_{i1} + \hat{\Psi}_{j1} - 2\hat{\Psi}_{ij}$$

Because $\hat{\Psi}$ is positive definite, the matrix $\hat{\Gamma}$ is necessarily conditionally negative definite.

Step 6. Plot $\hat{\gamma}_{ij}$s against $\|s_i - s_j\|$. This may not be a smooth function. A priori, it is assumed that the variogram function should be a reasonably smooth function of the distances. To obtain such a function, steps 2–5 are repeated until the resultant plot is visually smooth. This can be done automatically by putting a constraint that the difference $\sup_{i,j}|\hat{\gamma}_{ij} - \tilde{\gamma}_{ij}|$ is less than a prespecified tolerance. (Recall that $\tilde{\gamma}_{ij}$ denote the smooth estimates after fitting the smoothing spline.) The smoother may be constrained to be a monotone function of the distance.

At the end of this cycle, we have a collection $(\|s_i - s_j\|, \tilde{\gamma}_{ij})$ such that the matrix $\tilde{\Gamma}$, obtained from $\tilde{\gamma}_{ij}$s, is conditionally negative definite. Along with this, we also have a spline interpolant function that is passing through the $\tilde{\gamma}_{ij}$s. The

key question is: Is this interpolant a conditionally negative definite function? A question similar to this one, was posed by Myers (1984) in his comment on Dunn (1982). "The conditional negative semidefiniteness (CNSD) property must be satisfied for all observed locations and all locations for which an estimate is desired." The following theorems can be used to prove the required result.

The following result is known for spline interpolants.

Result 1. Let f be a continuous function defined on a bounded range and have bounded variation. Suppose f is known only at finitely many points; that is, we are given $(x_i, f(x_i))$, $i = 1, 2, \ldots, l$ only. Then as n increases,

$$\sup |f(x) \quad f_s(x)| \to 0$$

where $f_s(x)$ is the spline interpolant of the data $(x_i, f(x_i))$, $i = 1, 2, \ldots, l$.

In other words, spline interpolant is arbitrarily close to the original function as long as l is large.

Result 2. Let $\gamma(\cdot)$ be a variogram function, that is, a conditionally negtative semidefinite (CNSD) function. Let it be known only at finitely many points, that is, we are given $(d_i, \gamma(d_i))$, $i = 1, 2, \ldots, l$. If l is large enough, the spline interpolant of $(d_i, \gamma(d_i))$ also is conditionally negative semidefinite.

Proof. Let $\gamma_s(d)$ denote the spline interpolant. From this previous result, we know that given $\epsilon > 0$, there exists an $l(\epsilon)$ such that for all $l > l(\epsilon)$,

$$\sup_d |\gamma(d) \quad \gamma_s(d)| < \epsilon$$

Let $a_1, \ldots, a_n$ be any n real numbers $s_1, s_2, \ldots, s_n$ be n locations. Let $\Sigma_i^n a_i = 0$ and $a^T a = 1$ and $d_{ij} = \|s_i - s_j\|$ $i, j = 1, 2, \ldots, n$ be the distances between locations. Because $\gamma(\cdot)$ is a valid variogram,

$$\sum_i \sum_j a_i a_j \gamma(d_{ij}) < 0$$

We also know that

$$\gamma(d_{ij}) - \epsilon < \gamma_s(d_{ij}) < \gamma(d_{ij}) + \epsilon$$

Hence

$$\sum_i \sum_j a_i a_j \gamma_s(d_{ij} < \sum_i \sum_j a_i a_j \gamma(d_{ij}) + \epsilon$$

Because ϵ is arbitrary, the right-hand side can be made less than or equal to zero (by selecting large enough l). This proves the result.

The following result seems intuitively true, a formal proof would be of interest.

Result 3. Let $s_1, \ldots, s_n$ be n locations. Let $d_{ij} = \|s_i - s_j\|$. Let C be an $n \times n$ matrix with $C = (c(d_{ij}))$. If C is a conditionally negative definite

matrix, then there exists at least one valid variogram function $\gamma(\cdot)$ such that

$$\gamma(d_{ij}) = c(d_{ij}) \qquad i, j = 1, 2, \ldots, n$$

at least as $n \to \infty$ such that $\{d_{ij}\}$s are dense in $\mathcal{R}$. In other words, as $n \to \infty$ corresponding to each conditionally negative definite matrix of dimension $n \times n$, there is a variogram model.

Combining these three results, one can argue that the interpolant obtained at the end of Step 6 is a valid variogram function. In the simulations as well as various data analysis, the interpolant was a valid variogram function.

To summarize the discussion: We obtain a moment estimator of the underlying variogram, smooth the moment estimator using a spline smoother. The spline smoother is evaluated at a finitely many points. These points are approximated by a collection of conditionally negative definite values (Steps 3–5). Then, these conditionally negative definite values are smoothed using a spline smoother. These steps are repeated until the spline smoother and the spline interpolator are identical. At this stage, theorems 1 and 2 which apply to an interpolant can be evoked to show that the resultant variogram, in fact, is conditionally negative definite.

On the other hand, Cherry (1994), Shapiro and Botha (1991), and Hall, Fisher, and Hoffman (1994) try to smooth the moment estimator using functions which are conditionally negative definite. For two-dimensional data, the basic functions used for this smoothing are Bessel functions, J_0, which are themselves infinite series. This class is difficult to treat mathematically and computationally especially when conditions of smoothness, monotonicity, etc., are imposed. Moreover, none of the studies include performance of their variogram estimator in terms of prediction and prediction error. Theoretical and computational comparisons between the two classes of approaches would be of interest.

SIMULATION STUDY

In order to study the practical efficacy of the suggested kriging procedure, a simulation study was conducted. The purpose of this simulation study was to study the performance of the nonparametric variogram estimator and also the effectiveness of the principal components based kriging neighborhood.

As described earlier, a variogram is used for the purpose of predicting an observation at a new, unsampled location. Thus, the performance of a variogram may be judged in terms of the accuracy of the prediction or, more generally, the coverage properties of the corresponding prediction intervals. By coverage properties of a prediction interval, we mean the percentage of times the true value is contained in the prediction interval. Of course, the coverage percentage is not the only feature of the prediction interval that is important; its length is also important. Ideally, one wants a short interval with good coverage.

In this paper, it is suggested that one may use only the first few principal components for obtaining the predictions. This obviously reduces the computational burden substantially. The usual strategy for reducing the computational burden, when faced with a large data, is to select a *geographical* neighborhood of the new location and base the predictions only on the sample within this region. Our approach emphasizes the *statistical* neighborhood which is a global neighborhood in the sense of Davis and Grivet (1984). The coverage properties of the prediction intervals based only on the first few principal components are studied. The point predictor obtained is unbiased; however the prediction error and hence the associated prediction interval is not the shortest. It is observed, based on the simulation results, that the difference between the optimal interval and this interval is ignorably small but the computational simplicity obtained is substantial.

The simulation study was conducted as follows.

Step 0. Let COUNT = 0, TRUE = 0, AVPE = 0

Step 1. Generate 65 observations on a regular grid under the exponential variogram model (Cressie, 1991, p. 61)

Step 2. Based on $z(s_1), \ldots, z(s_{64})$, predict $Z(s_{65})$. Let us denote this predicted value by $p(Z; s_{65})$.

Step 3. Calculate the prediction error [(Eq. (4)] associated with $p(Z; s_{65})$. Let us denote it by p.e. $(Z; s_{65})$.

Step 4. Calculate the prediction interval, namely

$$p(Z; s_{65}) \pm 2\sqrt{p.e.(Z; s_{65})}$$

Step 5. Check if $z(s_{65})$, the 65th observation belongs to the prediction interval. If it belongs to the interval, then

$$\text{COUNT} = \text{COUNT} + 1$$

Step 6. Calculate

$$\text{TRUEPE} = \text{TRUEPE} + [z(s_{65}) - p(Z; s_{65})]^2$$

$$\text{AVPE} = \text{AVPE} + p.e.(Z; s_{65})$$

Step 7. Repeat Steps 1–6, B number of times. We select $B = 225$

Step 8. Calculate

$$\text{Coverage probability} = \frac{\text{COUNT}}{B}$$

$$\text{True prediction error} = \frac{\text{TRUEPE}}{B}$$

$$\text{Average estimated prediction error} = \frac{\text{AVPE}}{B}$$

Similar quantities were calculated for data of size 145. First 144 observations were used to predict the 145th observation.

If the variogram estimator and the estimator of prediction error are good, then coverage probability is close to the nominal coverage of 95%. Moreover, "true prediction error" and "average estimated prediction error" also are similar. To check the correctness of the data generation algorithm, the true variogram was used to obtain $p(Z; s_{65})$ in Step 2 and p.e. $(Z; s_{65})$ in Step 3. The coverage probability as well as the average prediction error are close to the right quantities. This is illustrated in the first row titled "true variogram" of Tables 1 and 2. These illustrate what is known as the "optimal error." The second row consists of the results if the nonparametric variogram estimator was used to obtain $p(Z; s_{65})$ in Step 2 and p.e. $(Z; s_{65})$ in Step 3. These reflect what is

Table 1. Simulation Results with Exponential Variogram Model[a]

	Variability (%)	Coverage (%)	Avg. estimated prediction error	Actual prediction error
$C_0 = 1, C_1 = 1, C_3 = 1$				
True variogram		95.9	1.2791	1.2013
Nonparametric variogram	60%	92.3	1.2687	1.4861
	70%	89.0	1.2093	1.5775
	80%	88.8	1.1972	1.5942
	100%	86.6	1.1866	1.5852
$C_0 = 0, C_1 = 3, C_3 = 0.5$				
True variogram		96.4	0.8631	0.8299
Nonparametric variogram	60%	92.4	1.5757	1.7990
	70%	87.6	1.4565	1.9362
	80%	85.8	1.4115	1.9519
	100%	83.6	1.3945	1.9831
$C_0 = 1, C_1 = 3, C_3 = 2$				
True variogram		93.3	1.4076	1.4993
Nonparametric variogram	60%	80.0	1.4376	2.1436
	70%	73.5	1.3535	2.2770
	80%	72.3	1.2780	2.3363
	100%	70.3	1.2486	2.3690
$C_0 = 0, C_1 = 3, C_3 = 0.5$				
True variogram		95.2	0.7351	0.7381
Nonparametric variogram	60%	95.2	0.9720	0.9698
	70%	93.7	0.9569	0.9865
	80%	94.0	0.9555	0.9892
	100%	93.3	0.9540	0.9905

[a]Exponential variogram model is $\gamma(h) = C_0 + C_1 \exp\{-h/C_3\}$ for $h \geq 0$ where h is distance between locations. Under this model 225 datasets, each of size 65, were generated. 65th data point was predicted based on 64 observations. Percent coverage corresponds to percentage of times actual observation was covered by prediction interval.

Table 2. Simulation Results with Exponential Variogram Model[a]

	Variability (%)	Coverage (%)	Avg. estimated prediction error	Actual prediction error
$C_0 = 1, C_1 = 1, C_3 = 0.25$				
True variogram		98.6	0.6854	0.6196
Nonparametric variogram	60%	94.7	0.9480	0.9172
	70%	93.3	0.9087	0.9594
	80%	92.4	0.9006	0.9664
	100%	92.0	0.8984	0.9713
$C_0 = 1, C_1 = 3, C_3 = 1$				
True variogram		94.7	1.3237	1.2335
Nonparametric variogram	60%	78.2	1.4557	2.3786
	70%	70.7	1.3545	2.6223
	80%	67.1	1.2345	2.7382
	100%	67.1	1.2128	2.7134
$C_0 = 1, C_1 = 1, C_3 = 1$				
True variogram		95.6	1.2767	1.2649
Nonparametric variogram	60%	92.4	1.2530	1.4172
	70%	86.7	1.1717	1.5245
	80%	85.8	1.1602	1.5403
	100%	86.2	1.1554	1.5319
$C_0 = 0, C_1 = 3, C_3 = 0.5$				
True variogram		95.3	0.2445	0.2381
Nonparametric variogram	60%	91.1	1.5702	1.7757
	70%	83.6	1.3912	1.9320
	80%	79.1	1.2489	1.9834
	100%	75.1	1.1920	2.0292

[a]Exponential variogram model is $\gamma(h) = C_0 + C_1 \exp\{-h/C_3\}$ if $h \geq 0$ where h is distance between locations. Under this model 225 datasets, each of size 145, were generated. 145th data point was predicted based on 144 observations. Percent coverage corresponds to percentage of times actual observation was covered by prediction interval.

known as the "actual error." The "Percentage variability" column tells how many principal components were utilized to obtain the predicted values $p(Z; s_{65})$ and the corresponding prediction error p.e. $(Z; s_{65})$ using Equation (5). For example, 60% implies the number of components that explained 60% of the variability in the Ψ matrix. It does not correspond to 60% (i.e., 38 for $n = 64$) of the number of principal components. Usually first 7–10 (for $n = 64$) principal components explain 60–70% variability.

The qualitative conclusions of this study are:

(1) Usually the first few principal components (accounting for about 70% of the total variability) are enough to get a prediction interval which is close to the optimal interval.

(2) The estimated prediction error based on the same number of principal components, in general, is better reflective of the actual prediction error. This is the reason for getting good coverage properties for the prediction intervals based on the first few principal components. Thus, although these intervals are not the shortest, *the confidence statements based on them are reflective of the truth.* This probably is due to the result that estimation of the large eigenvalues and the corresponding eigenvectors tends to be numerically stable.

(3) The change in the prediction error and the actual point prediction is small after the first few principal components. Both seem to settle down at about the same number of principal components.

A REANALYSIS OF CRESSIE'S (1986) IRON-ORE DATA

This is a standard data set described and analyzed by Cressie (1986) and Zimmerman and Zimmerman (1991). This dataset consists of iron-ore measurements taken from an orebody in Australia. The data and their spatial locations were displayed in Cressie (1986). The spatial locations form an incomplete rectangular grid whose internodal spacing equals 50 meters. The residuals from median polish are used for kriging. Anisotropy in the N–S direction was noticed by Cressie (1986) and corrected by doubling the scale in that direction. Instead of fitting a parametric variogram (Zimmerman and Zimmerman, 1991), the nonparametric variogram estimator is used as suggested in this paper. The estimate of the variogram superimposed on the moment estimator is shown in Figure 1. This figure also shows the parametric variogram estimator obtained by Zimmerman and Zimmerman (1991). This estimator evidently is close to

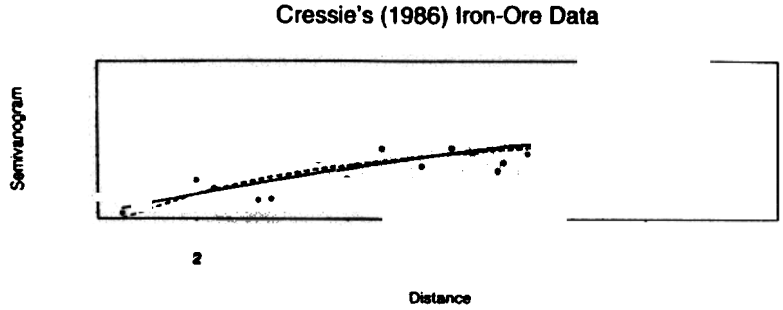

Figure 1. Estimated semivariogram for iron-ore residuals based on all distances less than 9.5525 regardless of direction. Weighted least-squares fit of parametric variogram (used by Zimmerman and Zimmerman, 1991) is shown by dotted line. Nonparametric variogram estimate is shown by solid line. Similarity of two variograms is striking.

theirs, except that this estimator does not use any parametric assumptions. The same locations are predicted as those predicted by Zimmerman and Zimmerman (1991). The results are displayed in Table 3. The conclusions are:

(1) The predictions obtained under the nonparametric variogram are similar to the one obtained by Cressie (1986) and Zimmerman and Zimmerman (1991).

(2) The prediction intervals given here are almost of the same length.

(3) If one uses the predictions and prediction intervals based on the first 12 principal components (explaining around 70% variability), they are extremely close to the optimal predictions and prediction intervals. Thus, a tenfold saving in computation is feasible, in addition to the advantage of assuming no parametric models.

The decision to select 12 principal components was based on the decrease in the prediction error as described previously. Graphically it is shown how the

Table 3. Predicted Values and Associated Prediction Interval at Two Locations for Cressie's (1986) iron-ore Data[a]

Location	Method	Predicted value	Prediction interval	Length of the prediction interval
	Nonparametric (12 PCs)			
	No monotonicity	0.3842	-4.6743,5.4426)	10.1169
	With monotonicity	0.3025	-4.8159,5.4209)	10.2368
	(All PCs)			
	No monotonicity	0.0839	(−5.0191,5.1871)	10.2062
	With monotonicity	0.9248	(−3.9987,5.8484)	9.8471
	Parametric			
	Zimmerman and Zimmerman (1991)	-0.2500	-5.2900,4.8000)	10.0900
	Nonparametric (12 PCs)			
	No monotonicity	.0954	(−4.1169,6.3078)	10.4247
	With monotonicity	.3497	(−3.8917,6.5911)	10.4828
	(All PCs)			
	No monotonicity	0.9759	-4.1372,6.0891)	10.2263
	With monotonicity	1.0945	-4.1071,6.2963)	10.4034
	Parametric			
	Zimmerman and Zimmerman (1991)	.0700	(−4.0300,6.1800)	10.2100

[a]See Zimmerman and Zimmerman, 1991 for detai

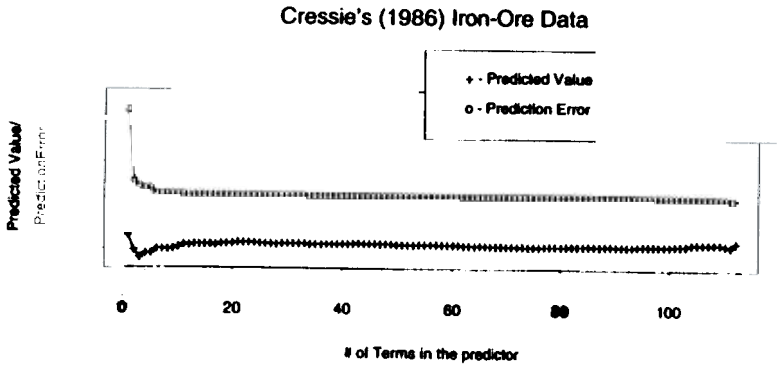

Figure 2. This plot shows predicted values and corresponding prediction errors against number of terms included in predictor. It is clear that inclusion of first few terms in predictor reduces prediction error close to optimal prediction error.

prediction error decreases with addition of principal components. In Figure 2, we display the change in the prediction error and the actual point predictor against the number of principal components. From this display, it is concluded that the first 12 principal components should be included in the prediction process because the changes after that are small. This also corresponds to about 70% explained variability. This graphical display proves to be a useful tool for determining the statistical kriging neighborhood.

DISCUSSION

Availability of model robust, automated, and computationally simple method for spatial prediction is important for environmental data analysis and decision making. The use of nonparametric variogram and principal components based kriging is proposed toward this goal. The simulation study and the analysis of a real-life dataset suggest that the methodology described here is useful in practice.

The idea of using "statistical," instead of a "geographical," neighborhood is applicable although a parametric model for variogram is used. A reasonable conjecture would be that it improves the estimation of the prediction error as well as the coverage properties of the prediction interval in a fashion similar to the one reported for nonparametric variogram estimator.

ACKNOWLEDGMENTS

This work was supported partially by EPA (CR-8200860-03) 1 and DOE (DE-FC07-94ID13317) grants to Prof. Goodman, Department of Biology, Montana State University, Bozeman, Montana. The author gratefully acknowledges comments by Noel Cressie, Daniel Goodman, and Mark Taper. Computer programs were written by Mr. Vendantham. The comments by the two anonymous referees has improved the original version of this paper. I thank them and those involved in the editorial process for their patience and encouragement.

REFERENCES

Armstrong, M., and Diamond, P., 1984, Testing variograms for positive definiteness: Math. Geology, v. 16, no. 4, p. 407-421.

Cherry, S., 1994, Nonparametric estimation of variogram: unpubl. doctoral dissertation, Montana State University, 143 p.

Christakos, G., 1984, On the problem of permissible covariance and variogram models: Water Resources Research, v. 20, no. 2, p. 251-265.

Cressie, N. A. C., 1986, Kriging nonstationary data; Jour. Am. Statis. Assoc., v. 81, no. 395, p. 625-634.

Cressie, N. A. C., 1991, Statistics for spatial data: John Wiley & Sons, New York, 900 p.

Davis, M. W., and Grivet, C., 1984, Kriging in a global neighborhood: Math. Geology, v. 16, no. 3, p. 249-265.

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R., 1975, Robust estimation and outlier detection with correlation coefficients: Biometrika, v. 62, no. 3, p. 531-545.

Dunn, M. R., 1982, Conditions for a variogram model to yield positive variances under restrictions: Math. Geology, v. 15, no. 4, p. 553-564.

Golub, G. H., and Van Loan, C. F., 1989, Matrix computations (2nd ed.): Johns Hopkins Univ. Press, Baltimore, Maryland, 642 p.

Hall, P., Fisher, N. I., and Hoffman, B., 1994, On the nonparametric estimation of covariance functions; Annals of Statistics, in press.

Jolliffe, I., 1986, Principal component analysis: Springer Verlag, New York, 271 p.

Journel, A. G., 1977, Kriging in terms of projections: Math. Geology, v. 9, no. 6, p. 563-586.

Journel, A. G., 1988, New distance measures: the route towards truly non-Gaussian geostatistics: Math. Geology, v. 20, no. 4, p. 459-475.

Journel, A. G., and Huijbregts, C. J., 1978. Mining geostatistics: Academic Press, London, 600 p.

Kacewicz, M., 1991, Solving the kriging problem by using the Gram-Schmidt orthogonalization: Math. Geology, v. 23, no. 1, p. 111-118.

Mead, A., 1992, Review of the developments of multidimensional scaling methods: The Statistician, v. 41, no. 1, p. 27-39.

Myers, D., 1984, Conditions for a variogram model to yield positive variances under restrictions: a comment: Math. Geology, v. 16, no. 4, p. 431-432.

Rousseeuw, P. J., and Molenberghs, G., 1993, Transformation of non positive semidefinite correlation matrices: Communication in Statistics-Theory and Methods, v. 22, no. 4, p. 965-984.

Schoenberg, I. J., 1938, Metric spaces and completely monotone functions; Annals Mathematics, v. 39, no. 4, p. 811-841.

Shapiro, A., and Botha, J. D., 1991, Variogram fitting with a general class of conditionally nonnegative definite functions: Computational Statistics and Data Analysis, v. 11, no. 1, p. 87-96.

Silverman, B. W., 1984, A fast and efficient cross-validation method for smoothing parameter choice in spline regression: Jour. Am. Statis. Assoc., v. 79, no. 387, p. 584-589.

Silverman, B. W., 1986, Density estimation for statistics and data analysis: Chapman and Hall, New York, 175 p.
Vecchia, A. V., 1992, A new method of prediction for spatial regression models with correlated errors: Jour. Roy. Statis. Soc., Ser. B, v. 54, no. 3, p. 813–830.
Wahba, G., 1990, Spline models for observational data: Soc. Industrial and Applied Mathematics, Philadelphia, 169 p.
Zimmerman, D., and Cressie, N. A. C., 1992, Mean squared prediction error in the spatial linear model with estimated covariance parameters: Ann. Inst. Statist. Math., v. 44, no. 1, p. 27–43.
Zimmerman, D., and Zimmerman, M. B., 1991, A monte-carlo comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors: Technometrics, v. 33, no. 1, p. 77–91.

Mathematical Geology, Vol. 27, No. 5, 1995

Letters to the Editor

Comments on Lerche (1993) and Liu and Lerche (1993)

Lerche has published two more erroneous papers on hydrocarbon chemical kinetic models (Lerche, 1993; Liu and Lerche, 1993) to confuse the nonexpert. The issue at hand concerns the validity of Lerche's so-called "bulk" kinetic model and whether it is nearly equivalent to the "parallel" kinetic model used by others. We use Lerche's "bulk" terminology for convenience, even though the model has less validity for bulk conversion than the "parallel" model. Briefly, the two models are defined by

$$\text{parallel} \quad x = \sum f_i \exp\left(\int k_i \, dt \right)$$

$$\text{bulk} \quad x = \exp\left(-\sum f_i \int k_i \, dt \right)$$

where x is the unreacted fraction, k_i is the rate constant for the ith reaction channel and f_i is the fraction of the total reaction assigned to that channel. The rate constant usually is assumed to follow the Arrhenius equation:

$$k_i = A \exp\left(-E_i/RT\right)$$

where a common frequency factor, A, is used for all reaction channels.

Lerche (1993) used series expansions to argue that "there is no functional difference" in the bulk and parallel models. An indirect series expansion comparison is unnecessary, however, because the equations easily are integrated numerically throughout the entire extent of conversion. The resulting curves for fraction reacted at two heating rates are given in Figures 1 and 2 for four example discrete activation energy distributions relevant to modeling organic maturation. It is obvious that the two models agree for the trivial situation of a single channel and that the difference between the two approaches become more substantial as the width of the distribution increases. The one and two kcal/mol spaced distributions are typical of the magnitude determined for marine source rocks (Tissot, Pelet, and Ungerer, 1987; Schaefer and others, 1990; Braun and others, 1991; Jarvie, 1991; Sundararaman, Merz, and Mann, 1992), and the difference

0882-8121/95/0700-0693$07.50/1