Journal of Applied Science & Computations
Vol.1.No.2. October 1994

STATISTICAL METHODS FOR INTERPOLATING SPATIAL DATA

Donald E. Myers
Department of Mathematics
University of Arizona
Tucson, AZ 85721

ABSTRACT

The problem of interpolating sparse irregularly located spatial data occurs in many areas of application. Whether explicitly modeled or not, the interpolation process reflects uncertainty and it is important to have some measure of reliability associated with each estimated or interpolated value. There are two general approaches to interpolation; one corresponds to choosing a class of functions with one or more unknown parameters and using the data to estimate those parameters. A second corresponds to estimation of the value at a non-data location for as many non-locations as desired. The latter will implicitly define an interpolating function. Statistical methods provide tools for quantifying the uncertainty whichever approach is used.

Geostatistical methods ,which arose out of applications in mining, are essentially regression estimators and which over the last twenty five years have been utilized for a wide variety of applications. These methods are closely related to other more deterministic methods such as splines and radial basis functions. The connection between the two is based on a weak form of positive definiteness for kernel functions. In geostatistics these arise as spatial correlation functions and are estimable from the data allowing adaptability to particular phenomena. A brief review of the nature of the problem is provided and in particular of the importance of non-point data. The basics of geostatistics are reviewed and the connections with other methods are outlined. The role of positive definiteness is shown. Extensions, practical aspects such as software and examples of specific applications are discussed.

Key Words: Geostatistics, Interpolation, Smoothing, Kriging, Variograms, Positive Definiteness, Splines, Radial Basis Functions, Spatial Averages, Stationarity

I. INTRODUCTION

Spatial data is encountered in many areas of application. These include environmental monitoring and assessment, hydology, atmospheric sciences, soil physics, plant pathology, mining and geosciences. There are at least two characteristics of spatial data that distinguish it from non-spatial data. First, each data value is associated with a location and secondly, the value is generally associated with a support, i.e., an area or volume. In some cases it may be better to think of such data as the values of a measure rather than as the values of a function. Very commonly the data value is a spatial average over the area or volume. In some applications this area or volume may be small enough so that the data value can be associated with a single point. These values can be viewed as determining a function but in general the function is unknown. This is one of the features that distinguishes the problem from the usual interpolation or approximation application. There are at least two forms of uncertainty that may be a part of the problem. The data is often noisy or incorporates errors (such as measurement errors) but more importantly the "function" to be interpolated is unknown because its values are known only at a finite number of points. The uncertainty pertaining to the function may be reduced in some instances where the function is known to belong to a certain class or has certain properties such as continuity or differentiability.

The problem can be described somewhat more precisely as follows. Let $x_1,...,x_n$ be points in k-dimensional Euclidean space, $V_1,...,V_n$ be volumes centered at $x_1,...,x_n$ respectively. Let $Z_{V_1},...,Z_{V_n}$ be the data values. Let W be another volume and Z_W be the unknown value. A principal objective is to estimate this unknown value. The points $x_1,...,x_n$, i.e, the volumes $V_1,...,V_n$ may or may not be contained in W. A typical application might be that the data correspond to the (average) concentrations of a pollutant in soil samples and the objective is to estimate the average concentration over a large area in order to decide about remediation. Alternatively suppose that the soil sample volumes are quite small and the data can be considered to be associated with points instead of volumes. The objective may still be to estimate an average value over an area or volume. Alternatively the objective may be to interpolate the values at each node of a regular grid and then use these values to produce a contour plot. A third form of uncertainty is introduced by the estimation/interpolation process and when the data locations are irregularly located the reliability of the contour plot will not be spatially uniform. This uncertainty is really related to the uncertainty pertaining to the unknown function. Objectives may include finding or estimating the maximal or minimal value of the unknown function in an area or volume. When the data is noisy or includes measurement error the objective may simply be one of smoothing, i.e., removal of the noise terms. It will be seen that this can be incorporated into the general technique to be considered.

Unlike many standard approximation techniques where asymptotic or limit properties are of prime interest, many spatial data applications involve a fixed small number of data locations. It may be quite expensive or too intrusive to collect more data. For example hydraulic conductivity (a hydrologic flow parameter) is typically measured by pumping tests. Each additional data value requires a new well. In the case where data values are the concentrations of volatile organics, the collection of soil or water samples may require special handling and the cost of laboratory analyses is relatively high. One of the objectives may be the design of a sampling plan that is optimal in an appropriate sense.

Spatial data may also be generated as the values of a known function but one which is difficult to evaluate or one which it is difficult to integrate or to find extrema. (Powell, 1985)

In selecting a methology there are a number of desirable characteristics. It should be relatively easy to apply, there should be software readily available or easily produced and it should not be too demanding in terms of computational power. It should be sufficiently flexible in order to be able to apply it to different kinds of phenomena. There should be some potential for incorporating subsidary information into the interpolation process. The methodology should not only produce estimated or interpolated values, it should also provide some measure of the reliability of those estimates. It would be desirable for the technique to be equally applicable to different dimensional spaces. It would be desirable for the method to work equally well with point data as with non-point data. It may also be desirable to be able to extend the method to vector valued data since the latter is often encountered in environmental applications.

TWO EXAMPLES

A. Estimation of a Spatial Average.

Let V be a region in k-dimensional space and Z(x) a function defined on this space. Suppose the objective is to estimate the integral of Z over V. If V can be described in an analytic form, i.e., the limits of integration can be determined in analytic form, and the function is known in analytic form, i.e., it is given by a finite number of "formulas" and the anti-derivatives are easily determined then ordinary calculus will suffice. Under slightly weaker conditions the value of the integral can be approximated numerically. The simplest approach is to partition the region into subregions, compute the value of the function at a point in each subregion and form the sum of the products of the volumes and the functional values (this is then converted to an average by dividing by the total volume). Using the continuity or differentiability of the function one can specify the sizes of the subregions in order to ensure a sufficiently good approximation. Note that the value of the function at a point inside each subregion could be interpreted as an estimate of the average value over the subregion. Let $S = \{Z(x) \mid x \varepsilon V\}$. Consider S to be the set of values of a random variable W and let the distribution of this random variable be determined by a uniform probability on V. That is, for any $A \subseteq V$, P(A) = volume (A)/volume (V). Then the average of Z over V is simply the expected value of the random variable W. This is a standard statistical problem, n points are chosen "at random" from within V, the value of Z determined at each of these points and the arithmetic average of these values is the desired estimate. Moreover one could construct a confidence interval, i.e., provide a measure of reliability of the resulting estimate.

There are at least four difficulties that may arise. (1) The reliability is directly related to the sample size, i.e., the number of data points, (2) each of the data points must be inside V, (3) one must be able to select the locations where the values of Z are to be determined and (4) the values should be "point" values and not averages over small regions. The last of these is relatively easy to deal with. If the data are average values over subareas or volumes the problem may change in very significant ways. If V can be partitioned into congruent disjoint subregions for which average values can be determined then the method described above can be modified. The other three are more troublesome. In many applications the data has already been collected and there may be little possibility of collecting more, finally there may be data available at locations outside V and it would be desirable to use it. These problems, as well as the methods, are part of the motivation for the ideas and results to be described in the main part of this paper.

B. Smoothing data.

When data is generated by measurements or by the analysis of physical samples there may be errors introduced into the data by these processes. One objective in analysing the data may be to "remove" the noise resulting from these errors, a second may be to estimate smoothed values at unsampled locations. Let Y(x) be the unknown function and Z(x) be the noisy data then the following model might be used

$$Z(x) = Y(x) + E(x) \tag{1}$$

where E(x) represents the noise term. E(x) is frequently modeled as a Gaussian random function with mean zero and variance σ^2. In addition E(x), E(x+h) are assumed to be uncorrelated. There are a number of methods for removing the noise term and/or interpolating Z(x) at a non-data point. One of the most common is to assume that Y(x) is a linear combination of known linearly independent functions (such as monomials in the coordinates of x, or sines and cosines). The unknown coefficients in the representation for Y(x) are estimated by least squares. Alternatively Y(x) may be estimated/interpolated with the use of smoothing splines, these will be described in more detail later.

II. METHODOLOGY AND RATIONALE

A. Introduction

Let D be a region of interest in k-space, all volumes and data locations are assumed included in this region. An additional characteristic that is often empirically observed in spatial data is that it is spatially correlated. Intuitively this may be described in two different ways. First, data values at locations that are close together are more alike or similar than those far apart. Secondly, data values at nearby locations are more informative for producing estimates or interpolated values than values at far away locations. This is similar to but not identical to continuity of the unknown function, continuity requires that values at locations close together be close but puts no restriction on the relationship between values at locations far apart. Differentiability implies that the function is at least locally linear. A smoothness condition is a more global restriction. There are two problems with focusing on continuity or differentiability. One is that the data may be insufficient to quantify the degree of continuity and secondly it will be especially difficult in the case of non-point data. As will be seen later however, the derivation described in the next section is equivalent to an assumption of some form of smoothness. A glossary of geostatistical terms is found in Myers (1991c)

B. The Random Field Model

One way to incorporate the uncertainty concerning the unknown function is to consider the data as a (non-random) sample from a random function defined on the region D. That is, the unknown function belongs to a class of functions (called realizations of the random function) that is described in probabilistic terms. Alternatively, let $Z_{V_1},\ldots,Z_{V_n}$, Z_V be the values of n+1 jointly distributed random variables. It is well-known that the unbiased minimum variance estimator of

Z_v given the data is the condidtional expectation. Furthermore it is well-known that if these random variables are jointly Normally distributed then the conditional expectation is a linear combination of the data and is an L^2 projection. Joint Normality may be too strong an assumption in many applications (although in some cases joint Log Normality may be a reasonable assumption) yet it is difficult to model a joint distribution using data alone. The general form of the model is given by

$$Z(x) = Y(x) + m(x) + E(x) \qquad (2)$$

where Y(x) is a spatially correlated random function , m(x) is a deterministic component and E(x) is the noise or error term. Several different special cases will be considered:

(a) E(x) is identically zero and m(x) is a constant.

(b) E(x) is identically zero and m(x) is expressible as a linear combination of known linearly independent functions (with unknown coefficients)

(a') E(x) has mean zero, is not spatially correlated nor correlated with Y(x), and m(x) is a constant.

(b') E(x) has mean zero, is not spatially correlated nor correlated with Y(x), and m(x) is expressible as a linear combination of known linearly independent functions (with unknown coefficients)

Initially we will consider (a) and then show how to extend the results to the remaining cases. Formulating the interpolation problem in the context of random functions first appeared in the work of Matern (1986) (an English version of the 1960 Swedish version) and Matheron (1965).

Stationarity and correlation

As a general concept, stationarity of a random function means that some statistical characteristic is not position dependent. There are at least two possibilities. The first, simply called Stationarity, requires that the joint probability distributions of the random function be translation invariant. Since it is difficult to model even univariate distributions using only data, this form of stationarity is not often used in connection with estimation of spatial averages or interpolation. The second alternative is to require that the first two moments be independent of position and with a form of translation invariance. The random function Y(x) is Second Order Stationary if

(i) $E\{Y(x)\} = m$ exists and does not depend on x, i.e., it is a constant

(ii) $E\{Y(x + h)Y(x)\} - m^2 = C(h)$ exists and does not depend on x, h is a vector

C(h) is the covariance of Y(x)

The random function is Intrinsic Stationary of order Zero if

(i') $E\{Y(x)\} = m$ exists and is a constant

(ii') $\gamma(h) = 0.5\mathrm{Var}\{Y(x + h) - Y(x)\}$ exists and does not depend on x

$\gamma(h)$ is called the Variogram[1] of Y(x)

If Y(x) is Second Order Stationary then the covariance and the variogram both exist and $\gamma(h) = C(0) - C(h)$. Note that covariances are always bounded but variograms need not be. Brownian motion provides an example of a random function that is not second order stationary and the variogram is unbounded.

C. Positive Definiteness

Positive Definite Functions play an important role in certain forms of interpolation and provide a link between statistical methods and deterministic methods. In addition a weak form of conditional positive definiteness is useful, we digress here to introduce those ideas. Let g(h) be a real valued symmetric function defined on a region in k-dimensional Euclidean space. Then g is (strictly) positive definite if

$$\sum\sum b_i b_j g(x_i - x_j) > 0 \qquad (3)$$

for any points $x_1,...,x_n$ and any real coefficients $b_1,...,b_n$. The covariance functions of second order stationary random functions are positive definite. Such functions are representable as Fourier Transforms of postive measures. Positive Definite functions are related in a natural way to inner products on linear spaces and hence to norms. The quadratic form given in (3) can be written as a matrix product B^TGB where B is the column vector of the coefficients and G is the matrix of covariance values, G is then a positive definite matrix. It is easily shown that Postive Definite functions are bounded by the value at 0 and asymptotically go to zero as h gets large.

[1] Originally $\gamma(h)$ was called the *semivariogram* because $2\gamma(h)$ was called the variogram. However only $\gamma(h)$ is actually needed in the development and application of geostatistics. Over a period of time many authors have renamed $\gamma(h)$ as the *variogram*, that practice is followed herein.

Definition 1 Let g(x,y)=g(y,x) be defined (and real valued) at each point of A x A , where A is a subset of Euclidean k-space. Let W = $\{h_0,....,h_p\}$ be a collection of functions defined at each point of A and linearly independent on A. Then g is said to be (strictly) positive definite on A with respect to W if

$$\sum\sum b_i b_j g(x_i,x_j) > 0 \qquad (4)$$

for all points $x_1,....,x_n$ in A and all weight vectors $\{b_1,...,b_n\}$ such that

$$\sum b_i h_j(x_i) = 0 \ ; j = 0,...,p \qquad (5)$$

There are two important special cases worth noting; (i) g(x,y) is a function of x-y, (ii) after an affine transformation on A, g is a function of x-y. If C(h) is Positive Definite then C(h)-C(0) is conditionally positive definite with respect to the set of functions consisting of just the constant function (it is common in this special case to simply use the term conditionally positive definite and omit the reference to W). Note that conditionally positive definite functions need not be bounded. The following function is conditionally positive definite but not positive definite

$$-\gamma(r) = -r^{\beta} , 0<\beta<2 \qquad (6)$$

Given a positive definite or conditionally positive definite function with one or more parameters then it is easy to construct others.

Lemma 1.Let $g(x,y;\beta_1,...,\beta_s)$ be conditionally positive definite (or positive definite) for all $\beta_1,...,\beta_s$ in H (H the parameter space). Let $\phi(\beta_1,...,\beta_s)$ be non-negative on H (contained in R^k) then

$$\int_H g(x,y;\beta_1,...,\beta_s)\ \phi(\beta_1,...,\beta_s)d\beta_1...d\beta_s \qquad (7)$$

is positive definite in the same sense as g provided the integral of ϕ over H is finite and the integral in eq (7) exists.

While eq(7) is written in the form of an integral, it includes a important special case namely when ϕ is non-zero at only a countable number of points and hence the integral reduces to a positive linear combination. In the case of a finte sum, such a construction is known in the geostatistical literature as a nested model.

When the function g depends only the magnitude of h and not on its direction, the function is said to be istropic otherwise anisotropic. When the anisotropy is removable by an affine transformation then the anisotropy is called geometric.

We return now to the simplest objective, given values of the function at a finite number of points, estimate its value at a non-data point.Let $x_1,...,x_n$ denote the data points and let $Z(x_1),...,Z(x_n)$ be the data values. In addition to the question of which form of stationarity to use, there is a related question of whether the expected value of Y(x) is assumed known or not. This affects the form of the estimator/interpolator and the associated equations.

D. Simple versus Ordinary

Suppose that $E\{Y(x)\} = 0$ and $E\{Z(x)\} = m$ where m is a known constant and Y(x) is second order stationary then the estimator is written in the form

$$Z^*_{SK}(x) = m + \sum_{i=1,n}\lambda_i(x)\,[Z(x_i) - m] \qquad (8)$$

This form is called the Simple (Kriging) Estimator and is identified by the subscript SK. This has an intuitive interpretation, namely that the estimate is a modification of the mean. The coefficients are determined by imposing two conditions on the estimator; (1) the estimator should be unbiased, (2) the variance of the error of estimation should be minimal. That is,

$$E\{Z^*_{SK}(x) - Z(x)\} = 0$$

$Var\{Z^*_{SK}(x) - Z(x)\}$ is minimal (with respect to the coefficients)

Note that under the assumptions on Y(x) that Z(x) is also second order stationary. Note also that the coefficients in the estimator depend on x although in most of the literature this is not explicitly shown and will be deleted in the following sections.

The variance of the error of estimation is expressable as a quadratic form in the coefficients and the covariances

$$Var\{Z^*_{SK}(x) - Z(x)\} = \sum_{i=1,n}\sum_{j=1,n}\lambda_i(x)\lambda_j(x)C(x_i - x_j) - 2\sum_{i=1,n}\lambda_i(x)C(x-x_i).$$

By taking derivatives with respect to each of the λ's and setting these equal to zero, a system of linear equations is obtained

$$\sum_{j=1,n}\lambda_j(x)C(x_i - x_j) = C(x-x_i) \qquad ; i = 1,n \qquad (9)$$

The estimator is unbiased irrespective of the values of the coefficients in the estimator. This form of the estimator and the required assumptions make it less than useful in many applications but it is important for at least two reasons. First, in the case of mulivariate Normality this estimator coincides with the conditional expectation and hence it is not only the optimal *linear* estimator it is *the* optimal estimator. Secondly, it is closely related to the idea of a reproducing kernel Hilbert space where the inner product is given by the covariance. Note that the invertibility of the coefficient matrix in eq(9) is assured by the positive definiteness of the covariance function. The minimized variance is given by

$$\sigma_{SK}{}^2 = \sum_{j=1,n} \lambda_j (x) C(x_j - x) \tag{10}$$

This variance does not depend explicitly on the data values but rather only on the covariance function and the spatial pattern of the data locations as well as the particular location whose value is being estimated.

Now consider somewhat weaker conditions on Y(x) and hence on Z(x). As before assume that E{Y(x)} = 0, E{Z(x)} = m with m an unknown constant and that the variogram of Y(x) (and of Z(x)) exists. Then the estimator is given in the form

$$Z^*_{OK}(x) = \sum_{i=1,n} \lambda_i(x)\, Z(x_i) \tag{11}$$

To obtain unbiasedness, a constraint is imposed on the λ's, namely

$$\sum_{i=1,n} \lambda_i(x) = 1 \tag{12}$$

This is a sufficient condition but not necessary. Using this condition the variance of the error of

$$\mathrm{Var}\{Z^*_{OK}(x) - Z(x)\} = -\sum_{i=1,n}\sum_{j=1,n} \lambda_i(x)\lambda_j(x)\gamma(x_i - x_j) + 2\sum_{i=1,n}\lambda_i(x)\gamma(x-x_i)$$

In order to minimize this variance subject to the constraint, a Lagrange multiplier μ is introduced prior to differentiation. Then the resulting equations are

$$\sum_{j=1,n} \lambda_j(x)\gamma(x_i - x_j) + \mu = \gamma(x-x_i) \qquad ; i = 1,n \tag{13}$$

$$\sum_{i=1,n} \lambda_i(x) = 1$$

and the minimized variance is given by

$$\sigma_{OK}{}^2 = \sum_{j=1,n} \lambda_j (x)\gamma(x_j - x) + \mu \tag{14}$$

Both of these estimators are the same as or similar to regressions, in fact Goldberg (1962) derived these results using that perspective. The significant applications however originated with G. Matheron and his students, particularly in mining and hydrology. Because Matheron's early work was based in part on the work of D. Krige, Matheron called this estimation technique "krigeage" which in English became "kriging". The various estimators are called .... kriging estimators and the minimized variance is called the kriging variance.

III. WHY POSITIVE DEFINITENESS?

It may be helpful to re-write the equations in eq(9) in matrix form

$$C\lambda = C_0 \tag{15}$$

The positive definiteness of the covariance function ensures that the coefficient matrix in eq(15) is invertible. However when the equations in (13) are written in matrix form

$$\begin{bmatrix} K & U \\ U^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} K_0 \\ 1 \end{bmatrix} \tag{16}$$

where U is a column of 1's, the invertibility of the coefficient matrix is not ensured by the positive definiteness of the variogram since the negative of the variogram is only conditionally positive definite. Although in general K is not invertible, the coefficient matrix is invertible because of the conditional positive definiteness. This was shown in a slightly different context by Micchelli (1986) and more explicitly by Myers (1988a), it is implicit in Matheron (1973). Using the form given in eq(16) the estimator in eq(11) can be written as

$$Z^*_{OK}(x) = [Z(x_1),\ldots..,Z(x_n),0][\lambda^T \ \mu]^T = \Sigma_{j=1,n} b_j \gamma(x_j - x) + a \tag{17}$$

where the coefficients in the summation are obtained from the system

$$\begin{bmatrix} K & U \\ U^T & 0 \end{bmatrix} \begin{bmatrix} B \\ a \end{bmatrix} = \begin{bmatrix} Z \\ 0 \end{bmatrix} \tag{18}$$

B being the column vector of the b's and Z being the column vector of the data values. This system has two important properties. First, the subsystem $KB + Ua = Z$ simply expresses the fact that the estimator is "exact". That is, if one estimates the value at a data point (and using the data value at that point) then the estimated value will be the data value. Secondly, if $\lim_{|h| \to \infty} \gamma(h) =$ constant, then the summation term goes to zero and the interpolating function is a constant, sufficiently far away from the data locations. This is because $U^T B = 0$. In fact quite independently of the derivation based on the random function model, by imposing exactness and the condition $U^TB = 0$ one obtains the system in eq(18) and the conditional positive definiteness is sufficient to ensure a unique solution. In the numerical analysis literature the interpolating function given by the system in (16) is known as a radial basis function although in most of the literature the variogram is assumed to be isotropic hence the name "radial". The form given in eq(17) is sufficiently general to include the thin plate spline which is usually derived in a totally different way. This connection was noted by Kimmeldorf and Wahba (1970), Wahba (1975) later made much more explicit by Matheron (1980, 1981a, 1981b). An elementary proof in one dimension is given by Watson (1984). Watson also points out the connections with the Green's Functions for certain differential operators. Dolph andWoodbury (1952) showed some of these connections in the 1-dimensional case. Positive Definiteness is also important because it provides the link to radial basis functions that were first introduced by Hardy (1971).

The positive definiteness of the covariance and the conditional positive definiteness of the variogram are important for another reason. The equations in eq(9) and in eq(13) are obtained by minimizing the variance of the error of estimation. Variances are supposed to be positive by definition and the positive definiteness properties of the covariance and variogram respectively ensure this. The non-negativity of the variance is also important to ensure that the minimized value is not negative because that would correspond to maximization (of the magnitude) instead.

A. An Abstract Formulation

Let H_1, H_2 be Hilbert spaces of functions on $R^k$and A a bounded linear operator from H_1 into H_2. Let $L_1,\ldots,L_n$ continuous linear functionals on H_1. For example, the linear functionals could be point evaluations at the points $x_1,\ldots, x_n$. Then the abstract spline is the function f^* in H_1 such that

(a). Af^* has minimum norm in H_2

(b). the $L_1f^*,\ldots,L_nf^*$ have prescribed values

The connection with the estimator/interpolators given above is constructed in several steps:

i. re-norm H_1 via A using the norm in H_2

ii. unless A is 1-1, the previous step will only produce a semi-norm, form the quotient space using the null space of this semi-norm and thus obtain a norm

iii. on the Hilbert space the norm defines an inner product

iv. the inner product is defined by a quadratic form which in turn is given by a positive definite function (or if a quotient space is used then the norm is induced by a conditionally positive definite function.

This is the construction used by Matheron (1980, 1981a, 1981b). The Simple Kriging equations are very similar to the representation obtained in a reproducing kernel Hilbert space and the estimator can be viewed as a projection onto the plane determined by the "data" vectors. In the case of Ordinary Kriging however, the projection is onto a hyperplane. An elementary discussion of this is given in Journel (1977). The abstract formulation and in fact the "kriging" formulation have the advantage that they are not dimension dependent, i.e., the dimension of the domain. However when the data locations are sparsely located in higher dimensional space there are still practical problems as noted by Foley (1987).

B. Ordinary versus Universal

Returning to the model given in (2), let E(x) be identically zero and let m(x) be a linear combination of monomials in the coordinates of x. Let these monomials be denoted by $f_0(x),\ldots,f_p(x)$ indexed in no particular order except that $f_0(x)$ is identically one. There is an interdependence between the dimension of the space and the number of such monomials of degree less than or equal to a specified integer. In one dimension there are 2, 3, 4 monomials of degree less than or equal to 1, 2, 3 respectively. In three dimensions, there are 10 monomials of degree no more than 2. Suppose further than Y(x) has a variogram. Given data $Z(x_1),\ldots.,Z(x_n)$, the objective is to estimate the value Z(x). The Universal Kriging (UK) estimator will again be of the form given in eq(11)

$$Z^*_{UK}(x) = \sum_{i=1,n} \lambda_i(x)\, Z(x_i) \qquad (19)$$

To obtain unbiasedness under these more general conditions however, multiple constraints are imposed on the λ's, namely

$$\sum_{i=1,n} \lambda_i(x) f_j(x_i) = f_j(x) \qquad ; j = 0,\ldots, p \qquad (20)$$

These are sufficient conditions but not necessary. Using these conditions the variance of the error of estimation is

$$\mathrm{Var}\{Z^*_{UK}(x) - Z(x)\} = -\sum_{i=1,n}\sum_{j=1,n} \lambda_i(x)\lambda_j(x)\gamma(x_i - x_j) + 2\sum_{i=1,n} \lambda_i(x)\gamma(x-x_i)$$

In order to minimize this variance subject to the constraints, p + 1 Lagrange multipliers are introduced prior to differentiation. Then the resulting equations become

$$\sum_{j=1,n}\lambda_j(x)\gamma(x_i - x_j) + \sum_{k=0,p}\mu_k f_k(x_i) = \gamma(x-x_i) \quad ; i = 1,n \tag{21}$$

$$\sum_{i=1,n}\lambda_i(x)f_k(x_i) = f_k(x) \qquad ; k = 0,\ldots,p$$

and the minimized variance is given by

$$\sigma_{UK}{}^2 = \sum_{j=1,n}\lambda_j(x)\gamma(x_j - x) + \sum_{k=0,p}\mu_k f_k(x) \tag{22}$$

The analogue of the estimator given in (15) will be

$$Z^*(x) = \sum_{j=1,n}b_j\gamma(x_j - x) + \sum_{k=0,p}a_k f_k(x) \tag{23}$$

When the variogram has a particular form known as "pure nugget effect"

$$\gamma(h) = \sigma^2, \ h \neq 0; \ \gamma(0) = 0$$

then eq(23) will coincide with Trend Surface Analysis (least squares fitting to a polynomial) except at the data points since the Trend Surface Analysis estimator is a smoother and is not exact.This was shown by Marcotte and David (1988). One of the differences between eq(17) and eq(23) is in how the function extrapolates the data. In eq(17) it simply uses the arithmetic average of the data, in eq(23) it uses a polynomial. The monomials could equally well be be replaced by any set of linearly independent functions although in Matheron's original development (1973) the monomials are important because the collection of monomials of degree ≤ s is closed under translations. The extended form of conditional positive definiteness given in Definition 1 is sufficient to ensure that the coefficient matrix in eq(21) is invertible.

C. Interpolation vs Extrapolation

There are at least three aspects of this issue. In general one would be very cautious about using an interpolator or estimator to estimate outside of the convex hull of the data locations. For example, the coefficients in a Trend Surface interpolator are determined only by the data and one can not separately incorporate information about extrapolation. In contrast the interpolator given by eq(23) implicitly allows incorporation of information about how extrapolation is performed. This is seen in part by examining what happens to the coefficients in eq (23) as the magnitudes of the vectors x_j - x increase. For a variogram with a sill it is easy to show that the first term in eq (23) goes to zero as the distance increases beyond the range. Hence only the second term affects the extrapolation. The same property holds in an asymptotic form for variograms without a sill. Since the uniqueness of the solution to the equations in eq(21) is ensured by the positive definiteness (with respect to the linearly independent functions) the choice as to how the interpolator extrapolates is somewhat arbitrary. This is not a characteristic of the Trend Surface estimator. A second point is whether the application of the estimator/interpolator must distinguish between interpolation and extrapolation and whether it has a measure of reliability that reflects the lack of information in the region of extrapolation. The kriging variance (for all three forms described above increases as the distance from the interpolation point to the nearest data location increases. Hence the reliability of the estimation at that point is significantly less. Finally in the case of the estimation of spatial averages, the "block" kriging estimator will use data locations inside and outside of the block. However, again the kriging variance will increase as the proportion of data locations outside the block increases. In many practical applications there will be few if any locations inside a specific block.

D.. Estimation/Interpolation vs Smoothing

Suppose now that in the model given in (2) that m(x) is a constant, Y has a variogram and E(x) is not identically zero. Let E(x) have mean zero and the covariance of E(x), E(x+h) be zero except when h = 0. Suppose further that the covariance of Y(x), E(x+h) is zero for any x , h. In this case the objective is to estimate Y(x) + m using the noisy data $Z(x_1),....,Z(x_n)$. Let σ^2 be the variance of the (spatially uncorrelated) error term E(x) and let $\gamma_Y(h)$ be the variogram of Y(x) (it is also the variogram of Y(x) + m). Note that the variogram of Z(x) and the variogram of Y(x) are not the same, for $h \neq 0, \gamma_Z(h) = \gamma_Y(h) + \sigma^2$. The estimator for Y(x) will be of the same form as eq (11) and eq(19).

$$Y^*(x) = \sum_{i=1,n} \lambda_i(x) Z(x_i) \qquad (11')$$

The same unbiasedness condition, eq(12), is utilized. However the variance of the error of estimation will have an extra term in it, namely

$$\sigma^2 \sum_{j=1,n}\lambda_j^2$$

and the system of equations will be slightly different:

$$\sum_{j=1,n}\lambda_j(x)\gamma_Y(x_i - x_j) + \lambda_i(x)\sigma^2 + \mu = \gamma(x-x_i) \qquad ; i = 1,n \qquad (24)$$

$$\sum_{i=1,n}\lambda_i(x) = 1$$

where the variogram of Y(x) is used. Of course the estimator will not be exact in that case. The smoothing spline can be written in this form. This connection has been pointed out by a number of people, for example Cressie (1990). The results are easily extendable to the case where m(x) is a linear combination of monomials. The Trend Surface estimator might be thought of as the extreme case of smoothing. That is, in the transition between interpolation and smoothing, it is at the smoothing end of the scale.

E. Estimation of Spatial Averages

Using one of the forms of the estimator above (with various kinds of assumptions) there are two possible ways to estimate a spatial average. One way would be to impose a grid over the region V and estimate the value at each grid point. For sufficiently small grid mesh the spatial integral is estimated by the arithmetic average of the estimated values at the grid points. Note that the data locations need not be inside the region V. One disadvantage is that it is difficult to determine the size of the grid mesh in order to obtain any given precision since the function Z is unknown and even its properties are unknown. Moreover the "error" is in two parts, the numerical integration error and the separate errors associated with the estimates at each of the grid points. Alternatively the spatial average could be directly estimated using an estimator of the same form as eq(11) and eq(19). Let Z_V be the average of Z over V, then

$$Z_V = (1/V)\int_V Z(y)dy \qquad (25)$$

$$Z_V^* = \sum_{i=1,n}\lambda_i Z(x_i) \qquad (26)$$

Let

$$\gamma(x,V) = (1/V)\int_V \gamma(x-y)dy \qquad (27)$$

and

$$\gamma(V,V) = (1/V^2)\int_V\int_V \gamma(u-y)dudy \tag{28}$$

Although written as single and double integrals eq(27) and eq(28) will actually be multiple integrals depending on the dimension of the space. The integral in eq(27) is the average value of the variogram between a point x and all possible points y in V (x might be inside or outside of V), this is sometimes called the "point-to-block average" of the variogram. The integral in eq(28) is the average of the variogram for all possible pairs of points in V. The equations for determining the coefficients in eq(26) can now be written in terms of the intergrals in eq(27) and eq(28). The equations are

$$\sum_{j=1,n}\lambda_j\gamma(x_i - x_j) + \mu = \gamma(x_i, V)) \qquad ; i = 1,n \tag{29}$$

$$\sum_{i=1,n}\lambda_i = 1$$

and the minimized variance is given by

$$\sigma_V{}^2 = \sum_{j=1,n}\lambda_j\gamma(x_j, V) + \mu - \gamma(V,V) \tag{30}$$

The estimator in eq(26) is commonly known as the "block kriging" estimator. The term "block" arises out of ore reserve estimation problems in mining. When an ore deposit is exploited it is common to determine physical "blocks" that will be classified as ore or waste. This is done on the basis of an average grade for the block, the grade is a proxy for the amount of "metal" in the block. The block kriging equations in (29) are derived in essentially the same way as for the Ordinary Kriging equations but with a few computational variations. If the unbiasedness condition in eq(29) is satisfied then

$$Z_V{}^* - Z_V = \sum_{i=1,n}\lambda_i Z(x_i) - (1/V)\int_V Z(y)dy = \sum_{i=1,n}\lambda_i (1/V)\int_V [Z(x_i) - Z(y)]dy \tag{31}$$

and it is easily seen that $E\{Z_V{}^* - Z_V\} = 0$. Moreover

$$[Z_V{}^* - Z_V]^2 = \sum_{j=1,n}\sum_{i=1,n}\lambda_j\lambda_i (1/V^2)\int_V\int_V [Z(x_i) - Z(u)]\,[Z(x_j) - Z(y)]dydu \tag{32}$$

Taking the expected value of both sides of eq(32) gives an expression for the estimation variance as a quadratic form in the λ's, the coefficients are terms of the form $\gamma(x_i - x_j)$, $\gamma(x_i,V)$ and $\gamma(V,V)$.

Again by incorporating Lagrange multpliers(s) and setting derivatives equal to zero the equations in (29) are obtained.

When implemented in software the integrals in eq(26) and eq(27) are normally computed numerically, in this case however the function (i.e., the variogram) is known. Hence error bounds are easily determined.

F. Equivalent and Alternative Methods

In the previous section it was shown that a spatial average could be directly estimated (albeit by an estimator of the same form as a point estimator). It can also be shown that there is a direct relationship between point estimation and the direct estimation of spatial averages. Suppose that a regular grid is superimposed on V with grid points $y_{st\cdot}$ each grid point is considered as the center of a small block v_{st}. These small blocks are assumed to be congruent and have common size, i.e., the volume is the same and will be denoted by v. The spatial average could then be estimated by a simple form of numerical integration

$$Z_V = (1/V)\int_V Z(y)dy \approx (1/V) \sum_{s=1,r} \sum_{t=1,q} Z(y_{st\cdot})v_{st\cdot} = (1/rq) \sum_{s=1,r} \sum_{t=1,q} Z(y_{st\cdot}) \tag{33}$$

However the $Z(y_{st\cdot})$ are not known and must be estimated. Then we have

$$Z_V \approx (1/rq) \sum_{s=1,r} \sum_{t=1,q} [\sum_{i=1,n} \lambda_i(y_{st\cdot})Z(x_i)] = \sum_{i=1,n} [(1/rq) \sum_{s=1,r} \sum_{t=1,q} \lambda_i(y_{st})]Z(x_i) \tag{34}$$

Now consider the kriging equations used to determine the coefficents in the estimators for the $Z(y_{st})$. Sum on both sides with respect to s, t.

$$[(1/rq) \sum_{s=1,r} \sum_{t=1,q} \sum_{i=1,n} \lambda_i(y_{st\cdot})\gamma(x_i - x_j) + [(1/rq) \sum_{s=1,r} \sum_{t=1,q} \mu st$$

$$= [(1/rq) \sum_{s=1,r} \sum_{t=1,q} \gamma(x_j, y_{st\cdot}) \tag{35}$$

Now simply interchange the order of summation in the first term on the left, rename the second term as μ and observe that the term on the right is a simple form of numerical integration applied to $\gamma(x_i,V)$. Hence in the limit the results are the same. It is more difficult to try to combine the separate point kriging variances into a block kriging variance.

Now consider approximating Z_V another way. Using the representation given in eq(17)

$$Z_V = (1/V)\int_V Z(y)dy \approx (1/V)\int_V Z^*(y)dy = (1/V)\int_V [\sum_{j=1,n} b_j \gamma(x_j - y) + a]dy$$

$$= \sum_{j=1,n} b_j \gamma(x_j, V) + a \tag{36}$$

Reversing the connection used in obtaining eq(17) we obtain the block kriging estimator. In this case then, integrating the interpolator is exactly the same as direct estimation. This is not a property of all interpolators and for most there is no counterpart of direct estimation.

G. Within and Between Block Variances.

The decomposition of variances plays an important role in many parts of statistics. There is an analogous idea connected with the estimation of spatial averages. Let V be a "large" block and suppose that it is represented as the union of disjoint, congruent "small" blocks v_{st}. There are three "variances" of interest. The first is the quantifies the variability of (point) values in V relative to the spatial average, this is given by

$$S^2(0,V) = (1/V)\int_V [Z(y)-Z_V]^2dy \tag{37}$$

For each "small" block there is a corresponding variance

$$S^2(0,v_{st}) = (1/v_{st}.)\int_{v_{st}} [Z(y)-Zv_{st}]^2dy \tag{38}$$

and finally there is the variability of the "small" block averages relative to the average over the "large " block.

$$S^2(v,V) = [(1/rq) \sum_{s=1,r} \sum_{t=1,q}[Zv_{st}.-Z_V]^2 \tag{39}$$

The integral in eq(37) can be re-written as a sum of integrals over the respective v_{st}'s. Then inside the brackets in eq(37), simply add and subtract Zv_{st} . Finally form the sum on the $S^2(0,v_{st})$ terms. This will represent $S^2(0,V)$ in terms of $S^2(v,V)$ and the sum on the $S^2(0,v_{st})$ terms. Now apply the expected value and one obtains

$$\gamma(0,V) = \gamma(0,v) + \gamma(v,V) \tag{40}$$

The first term is the average value of the variogram over all pairs of points in V, the second term is the average value of the variogram for pairs of points within one of the "small" blocks and the last term is the average value of the variogram between pairs of points, one in a randomly selected "small" block and one in the block V. For an example of an experimental verification of this result see Miesch (1975). Both in mining and in environmental remediation it is necessary to select a block size for decision making, i.e.. the size of a block to process as ore or waste, the size of a section of soil to remediate or to leave alone. In both cases the size of the decision block will be affected by the size of the equipment to be used and also by the acceptable decision error probabilities. The relationship in eq(40) is useful for planning in these situations, Parker (1979), Zhang et al (1990)

H. Kriging Variances and their use.

The kriging equations (Simple, Ordinary, Universal) are all derived by minimizing the variance of the error of estimation. Then this minimized variance can be computed using the solution of the kriging equations. Originally this was emphasized as a measure of the reliability of the estimates, i.e., of the estimation process. However this minimized variance (kriging variance) has a feature that is both an advantage and a disadvantage. The kriging variance does not depend on the data values except in an indirect way. For a known or specified variogram model, the kriging variance is a function of the sample location pattern and the spatial relationship of the location to be estimated to the data locations. Moreover the kriging variance is changeable by multiplying the variogram by a positive constant, the estimated data value is not affected by this operation however. This means that it is not quite the kind of variance that is useful for producing confidence intervals. One could use the kriging variance to compare the reliability of estimates at different locations (using the same variogram model). For example, if the estimates (and perhaps also the data) are used to generage a contour map then it is useful to also generate a contour map of the kriging standard deviations. Dowd (1992) reviews the more recent techniques for generating confidence intervals that do not suffer from the same defect.

Since the kriging variance(s) do not require a priori knowledge of the data, the kriging variance can be used to rank order sampling plans by optimizing the kriging variance. This has been promoted by a number of authors.

I. Two Special Cases

The Nugget variogram model corresponds to no spatial correlation. As was noted earlier, when a Nugget model is used with Universal Kriging, the results are the same as those obtained from a Trend Surface except at the data points. It is useful to consider a simpler version of this. Consider first the problem of estimating the value at a point x_0. By substituting the nugget model into eq(13) one obtains a simple solution, namely all the λ's are $1/n$ and $\mu = (1/n)\sigma^2$. That is, the estimator reduces to simply the arithmetic average of the data values and the kriging variance is $(1 + 1/n)\sigma^2$. Since all the data locations have the same weight, it is only the number of locations that affects the kriging variance.

Now suppose that instead of estimating the value at a point, the objective is to estimate a spatial average. Substituting the nugget model into eq(29), the solution is exactly the same and the estimator is identical but now the kriging variance is $(1/n)\sigma^2$. While intuitively point estimation might be thought of as the limiting case of block estimation, these results suggest that there is a discontinuity.

IV. MODELING SPATIAL CORRELATION

All of the preceding discussion assumes that the variogram (or the covariance function) is known. In practice this will not be the case and when the data is considered as a non-random sample from one realization of the random function it is not possible to "test" for stationarity. The potential for using the data to estimate and model the variogram is both an advantage and a disadvantage. It is an advantage in that the data can be used to customize the interpolating function and in that process auxillary non-quantitative information can be incorporated. It is a disadvantage in that it is not a completely automatic procedure and has a degree of subjectivity to it. Estimation and modeling of the variogram has a number of advantages over direct estimation and modeling of the covariance function, particularly in two or higher dimensions and when the data locations are not on a regular grid. First of all estimation of the variogram does not require separate estimation of the mean of Z (if the mean is a constant) whereas that is necessary for the covariance. Note that in the various systems of equations (11), (15), (17), (23) and (26), the left hand side of the system incorporates the spatial correlation of Z for pairs of data locations whereas the right hand side incorporates the spatial correlations between data locations and the point where an estimate is desired. On the left hand side estimated values might be used but the right hand side requires the use of a theoretical model for the variogram.

Estimation and modeling of the variogram can be approached in at least two different ways. Since the variogram is a function and must satisfy certain positive definiteness conditions, one could use some form of function fitting such as Maximum Likelihood or one could estimate the values of the variogram at a finite number of points and then "fit" a variogram to the estimated values. Both of these are used in practice usually in some combination. It is perhaps easier to focus on the second approach first. The most obvious choice for an estimator of values of the variogram is a sample variogram. With only a finite number of data locations, and hence only a finite number of pairings of points with only a finite number of interlocation vectors, the variogram can be estimated at only a finite number number of points.. Inasmuch as the variogram might exhibit an anisotropy it is necessary that the estimator provide information to allow determination of the directional dependence. The general form of the sample or experimental variogram is

$$\gamma^*(r,\theta) = (1/2N(r,\theta)) \sum[Z(x+h) - Z(x)]^2 \qquad (41)$$

where the sum is taken over all pairs of points x + h, x such that $r - \varepsilon/2 \leq |h| \leq r + \varepsilon/2$ and $\theta - \delta/2 \leq \arg h \leq \theta + \delta/2$. $N(r,\theta)$ is the number of pairs of locations satisfying these conditions. The interval $r - \varepsilon/2 \leq |h| \leq r + \varepsilon/2$ is called a distance class and is necessary because the data locations are generally not on a regular grid hence for any one distance there will be few if any pairs. The interval $\theta - \delta/2 \leq \arg h \leq \theta + \delta/2$ is called an angle window and is used because for any given direction and any given distance there will be few if any pairs. This does result in a smoothing and in fact some smoothing is desirable. Note that the sample variogram corresponds to a spatial average whereas the true variogram corresponds to an ensemble (probabilistic) average. This means that some form of ergodicity is implied otherwise one can not estimate the variogram using data from only one realization of the random function.

Typically the sample variogram is computed for a sequence of values of r of the form $\varepsilon/2$, $3/2\varepsilon$, ..., $(2n+1)/2\varepsilon$ with $\theta = 0, 45, 90$ and 135, $\delta = 45$. The particular case of $\delta = 180$ gives what is called the omni-directional experimental variogram. These experimental variograms separately provide information on how the variogram changes with separation distance and collectively they provide information as to whether the variogram is directionally dependent. In general the number of pairs is small for short separation distances, increases as the distance increases and then decreases again when the distance is approximately half the maximum separation distance. Usually the experimental variogram is not computed for distances larger than half the diameter of the region where the data is located. Some attention has been given to the question of optimal sampling designs for variogram estimations. As shown in Warrick and Myers (1987) however the optimal plans are not very practical. Norris (1991) has shown that because of the interdependence between pairs, the effective number of pairs is considerably less than the actual. Optimality for variogram estimation is quite different from optimality for kriging. In practice however the same data set is usually used for both and is probably not optimal for either. There are perhaps two principal difficulties with the sample variogram. First, it is an average and it is well-known that averages are not robust to outliers. Secondly, it is difficult to determine the distribution of the sample variogram. Davis and Borgman (1979, 1982) give general asymptotic results and exact results under very restrictive conditions. There are several other estimators than have been proposed but the experimental variogram is perhaps still the most common. Zimmerman and Zimmerman (1992) provide an analysis of the how various variogram estimators compare.

The computed values of the experimental variograms and their plots are used in several ways. Since the omni-directional variogram ($\delta = 180$) will generate the most number of pairs per plotted point, this one is usually examined first. There are three characteristics of particular interest. Is there a jump discontinuity at the origin? If so, the magnitude is the coefficient in a nugget effect component. Does the graph appear to increase to a maximum and then remain approximately constant? If so, the the magnitude of this maximum value (less the nugget component) is the sill of a second variogram component and the distance at which the leveling off takes place is the range. These graphical characteristics correspond to parameters in the models listed in the next section. Note that the experimental variograms may be relatively noisy and one does not simply fit a curve through the points. The experimental variograms for different directions are compared to determine whether an anisotropic model will be used, in particular does the range of the variogram change with direction. If the sample variogram exhibits quadratic growth this is usually taken as evidence that there is a non-stationarity. This is discussed in the next section. The mechanics of modeling variograms is discussed in greater detail in Myers (1991a, 1991c)

A. Valid Models

While the Bochner Theorem provides a theoretical test for positive definiteness and Matheron (1973) gives an extension of this theorem to conditionally positive definite functions, in practice it is non-trivial to test a function for either form of positivity. Hence a form of Lemma 1 is usually used. Using a small number of functions known to satisfy the conditions, positive linear combinations are constructed. In general the parameters in these models have a geometric or physical interpretation which aids in their utilization. The chosen model then is constructed as positive linear combination of valid models. Note that any valid model will produce an unbiased, minimum variance, linear estimator (BLUE). The minimum variance is computed using the variogram hence one can not choose between models by comparing kriging variances. Moreover the weights in the kriging estimators (Ordinary, Simple, Universal) are not affected by multiplying the variogram by a positive constant although the kriging variance will increase by that factor. Hence the kriging variances can be artificially increased or decreased without changing the estimated values.

In addition to the theoretical problem of ensuring that the variogram has the appropriate positive definiteness property there is the practical problem of choosing/fitting such functions. In practice this is done by the use of "nested models", i.e., postive linear combinations of known valid models. Moreover these are usually chosen from a relatively small list but because each component in the linear combination has one or more parameters, relatively complex models are easily constructed. The following is a list of models commonly used and commonly found in geostatistical software packages. Representing the variogram as a postive linear combination is analogous to representing the random function as a positive linear combination of uncorrelated random functions.

EXAMPLES (isotropic)

Spherical

$$\gamma(r) = \begin{cases} C_1\,\{1.5(r/a) - .5(r/a)3\}, & 0<r<a \\ C_1, & a < r \end{cases} \tag{42}$$

Nugget

$$\gamma(0) = 0 \qquad (43)$$

$$\gamma(r) = C_0 \text{ if } r > 0$$

Exponential

$$\gamma(r) \ = C_1\{1 - \exp(-r/a)\} \qquad (44)$$

Gaussian

$$\gamma(r) \ = C_1\{1 - \exp(-(r/a)^2)\} \qquad (45)$$

In all of the preceding models, C_1 is the sill. Each of these models corresponds to a random function with a finite variance and the sill is the variance. Note that if the variogram is modeled as a positive linear combination of the above model types then "sill" is used in two senses; first it refers to a parameter in each of the individual components and secondly it refers to the sum of the sills of the components. Sometimes these models are written with the nugget component incorporated but it is preferable to separate them, the software packages will require inputting the "sills" of the nugget term and as well as the sills of the other terms as separate parameters. The parameter *a* is either the range of dependence (in the case of the Spherical model) or is related to the range. It is also referred to as the range of "influence". Neither the Exponential nor the Gaussian have a true range but both have an effective range, the distance R at which $\gamma(R) = .95C_1$.

Power

$$\gamma(r) = \ C_1\, r^{\beta}, \quad 0 < \beta < 2 \qquad (46)$$

The Power model corresponds to a random function that does not have finite variance. When $\beta = 1$, it is called the Linear model. The Linear model has an important property, when used for kriging the coefficient C_1 does not affect the estimated values. Hence some software packages implement a kriging option using a Linear model and do not provide for any variogram estimation steps. The user should be aware that this is only a special case of kriging.

The list of models above is not exhaustive but includes the ones commonly found in geostatistical software packages along with a provision for nested models.

Note that if the smoothing form of kriging is used as in eq(11') and eq(24) then one must distinguish between modeling the variogram of Y(x) and modeling the variogram of Z(x). The nugget component of the variogram of Z(x) will include the variance of the error term E(x) but it may also include an amount reflecting short range variability nor attibutable to "noise". That is, the variogram of Y(x) may still include a nugget component.

B. Non-Stationarities

The experimental variogram is an unbiased estimator of the variogram if Z(x) has a constant mean. If the mean is not constant then the expected value of the experimental variogram will differ from the variogram by an additional term which is quadratic or higher order growth in h. Hence if the experimental variogram appears to show quadratic or higher order growth rate then (1) there are no valid models that would fit and (2) this growth is perhaps evidence of a non-stationarity. This leads to a circular problem. The optimal estimation of this non-constant mean requires knowing the variogram but the variogram is not easily estimated and model in the presence of a non-stationarity. Several solutions have been proposed and used, none is the full solution. If the data locations are on a regular grid then median polish is often an adequate estimator of the non-constant mean, this is subtracted from each data value to obtain residuals which are then used to estimate and model the variogram. See Cressie (1986) for an example of the use of median polish. Alternatively one can fit a low degree polynomial to the data using least squares, this is sub-optimal and can result in a bias in the experimental variogram. Some authors have used an iterative process. The non-stationarity is not a problem in the kriging step, only in the variogram estimation process. That is, by using relatively small search neighborhoods only a form of "local" stationarity is needed in order to use Ordinary Kriging. Alternatively one can use Universal Kriging where only the order of the drift is needed. Note however that if universal kriging is used then more data is required in the search neighborhood for any one estimation. Journel and Rossi(1989) have provided comparisons between using "local" stationarity with Ordinary Kriging vs Universal Kriging. A general discussion of stationarity in geostatistics is found in Myers (1989a).

C. Cross-Validation

The exactness of the kriging estimators can be used in a scheme to compare possible variogram models. For each choice of a variogram model the data locations are "jackknifed". That is, one at a time, each data location is temporarily suppressed and the value at that location estimated using only the remaining data. Note that in addition to choosing or fitting a variogram, the user must also choose the search neighborhood, the minimum and maximum number of data locations to be used for each estimation. For each data location then there are three values, the observed data value, the estimated value and the kriging variance (i.e., standard deviation). If the variogram model is "close" to the true but unknown model then collectively the observed data values should be close to the estimated values. However, "close" can be measured in several ways.

There are at least six statistics that can be used to evaluate a variogram model and its parameters. These include the mean error, the mean squared error,correlation of predicted vs observed and predicted vs prediction error. Because the derivations of the kriging equations (in the various forms) are not explicitly dependent on any distributional assumptions, it is difficult to construct statistical tests in the usual sense but some information is known about each of the statistics. The statistics can be written as

(i) $T_1 = (1/n)\sum[Z(x_i)-Z^*(x_i)]$

(ii) $T_2 = (1/n)\sum[Z(x_i)-Z^*(x_i)]^2$

(ii') $T_3 = (1/n)\sum\{[Z(x_i)-Z^*(x_i)]/\sigma\}^2$

(iii) $T_4 = \mathrm{Corr}\{[Z(x_i)-Z^*(x_i)]/\sigma, Z^*(x_i)\}$

(iv) $T_5 = \mathrm{Corr}\{Z(x_i), Z^*(x_i)\}$

Since $E\{T_1\} = 0$ the experimental value should be close to zero. Likewise since $E\{T_3\} = 1$, that experimental value should be close to one. The expected values of $E\{T_4\}$, $E\{T_5\}$ are more complicated since they depend on the Lagrange multipliers but they are "approximately" 0,1 respectively. The scatter plots of observed vs estimated and estimated vs error of estimation are also useful as are the histogram of the errors and the histogram of normalized errors. The latter may be useful in identifying anomalous data values. A data value could be anomalous in at least two different senses. First, it might be an outlier with respect to the distribution of data values and secondly it be an outlier spatially, i.e., significantly different from data values at nearby locations. Cross-validation is not a tool to determine the optimal variogram model but rather as a tool to compare several variogram models and/or a model with several different choices of parameters. The cross-validation statistics may be affected by other user choices such as the size of the search neighborhood and the maximum number of data locations used for each estimation. Some of the statistics are more sensitive to changes in the variogram type or the variogram parameters than are others. Note that cross validation used to evaluate variogram models and parameters is not the same as the use of cross validation to optimize the choice of the smoothing parameter for a smoothing spline. An extension of the use of cross validation to the selection of radial basis functions is given in Myers (1993)

D. Regularization:Non-point Data.

All of the preceding discussion assumes that the data used for interpolation and the data used for variogram estimation is "point" data. While nearly all data is actually non-point data it can often be considered as point data. For example, when the intersample location distances are large compared to the physical size of the sample there will be little affect resulting from the use of the non-point data. However there are many applications where the data is non-point data and the sample support is non-trivial. If data of the form $Z(x_1),\ldots.,Z(x_n)$ is replaced by data of the form Z_{v1}, .., Z_{vn} where the latter are spatial averages over volumes v1,....,vn then the kriging equations are easily modified to accomodate such data. It is only necessary to replace the $\gamma(x_i, x_j)$ terms by $\gamma(vi,vj)$ where the latter are average values of the variogram. However the problem is not so simple with respect to estimating and modeling the variogram. Suppose that the region of interest is partitioned into disjoint, congruent rectangular blocks. Label each such block by its center point, i.e., v_x. Let v be such a block centered at the origin. Then $v_x = \{y \mid y = x + u, u \varepsilon v\}$. That is, v_x is the block v translated by the vector x. Let Z_x, Z_{x+h} be the average values of Z over v_x , v_{x+h}. Consider then

$$\gamma_v(h) = 0.5\mathrm{Var}\{ Z_x - Z_{x+h}\} \qquad (47)$$

Under assumption (i) for the existence of variograms $E\{Z(x) - Z(x + h)\} = 0$ and hence eq(47) can be re-written as

$$\gamma_v(h) = 0.5E\{[Z_x - Z_{x+h}]^2\} = 0.5E\{[(1/v)\int_v[Z(x + u) - Z(x + u + h)]du]^2\} \qquad (48)$$

or

$$\begin{aligned}\gamma_v(h) &= 0.5E\{(1/v^2)\int_v\int_v[Z(x + u) - Z(x + u + h)]\,[Z(x + t) - Z(x + t + h)]dudt]\\ &=(1/v^2)\int_v\int_v \gamma(u-t+h)dudt \; -(1/v^2)\int_v\int_v\gamma(u-t)dudt \qquad (49)\end{aligned}$$

Theoretically then one effect of using non-point data is that the sill will be reduced. An additional characteristic is seen in the sample variogram, the shortest distance for which there is a plotted value will increase. This can make a non-nugget variogram look like a pure nugget when non-point data is used.

III. DEVELOPMENTS AND EXTENSIONS

A. The Vector Valued Case:Cokriging

In many applications there are multiple variables of interest and these are not only separately spatially correlated, they are also intercorrelated. In such cases it may be useful to utilize the data for all variables in estimating or interpolating any one of the variables. Let $Z(x) = [Z_1(x),\ldots,Z_m(x)]$ be a vector valued function. Then the analogue of the estimator in eq(11) is

$$Z^*(x) = \Sigma_{i=1,n} Z(x_i)\Gamma_i \qquad (50)$$

where the Γ's are m x m weight matrices. The unbiasedness conditions becomes

$$\Sigma_{i=1,n}\Gamma_i = I$$

There are at least three choices for functions to quantify the spatial intervariable correlation, the cross covariance, the cross variogram and the pseudo cross variogram. These are

$$C_{ij}(h) = \mathrm{Cov}\{Z_i(x+h), Z_j(x)\} \qquad (51)$$

$$\gamma_{ij}(h) = .5\mathrm{Cov}\{Z_i(x+h)-Z_i(x), Z_j(x+h)-Z_j(x)\} \qquad (52)$$

$$g_{ij}(h) = .5E\{[Z_i(x+h) - Z_j(x)]^2\} - .5[m_i - m_j]^2 \qquad (53)$$

When i = j, eq(51) becomes the covariance function, eq(52 and eq(53) become the usual variogram. While eq(52) is always an even function, eq(51) and eq(53) may not be. Hence eq(53) is more general than eq(52). One then constructs a matrix valued function whose diagonal entries are variograms (respectively covariances) and the off-diagonal entries are cross-variograms or pseudo-cross variograms (respectively cross covariances). The sum of the estimation variances can be expressed as the trace of an expression comparable to that obtained in the univariate case. Then a system of linear equations is obtained comparable to that for the univariate case but where all entries are matrices. The details are found in Myers (1982, 1983, 1984, 1988b, 1988c, 1989b, 1991b, 1991d, 1992). Modeling cross correlation functions is more difficult than modeling the variogram for a single variate. While the diagonal entries in the matrix function may be individually checked for the appropriate positive definiteness condition, this is not sufficient for the cross correlation components. This problem has not be totally solved in practical terms.

There is an important special case that should be noted. In some instances there is a primary variate of interest and one or more secondary variates. Obtaining data for the primary variate may be difficult, expensive or both whereas the secondary variate(s) less so. This is sometimes called the under-sampled case. This is similar to regressing the primary variate on one or more secondary variates but the estimator incorporates spatial correlation at a distance whereas ordinary regression does not.

Dowd(1992) provides an overview of more recent developments in geostatistics.

IV. NUMERICAL RESULTS

A. Software

At the present time, standard statistical packages such as SAS, SPSS, BMDP, SYSYTAT, etc. do not include routines for the algorithms discussed above although there are examples of individuals having written "procs" in SAS for these computations. Software is available, some in the public domain. The best known is a public domain package called GeoEas, released by the EPA, Las Vegas. This is available for the PC and the VAX (VMS). The use of this software is illustrated in Myers (1991c). A UNIX version of the GeoEas package is available by FTP from the author. Additional software can be found in various issues of the journal *Computers and Geosciences.* In comparing the results obtained using these algorithms one should consider several aspects. The user makes some decisions about parameters and these will effect the results. Secondly if the results are presented in the form of a contour plot then the contouring algorithm must also be considered. Packages that will contour irregularly spaced data include both an estimation/interpolation algorithm and a contouring algorithm. Since most contouring algorithms require data on a regular grid, the grid size used in the estimation/interpolation stage will affect the results.

B. Practical Aspects

As shown above the estimation/interpolation problem can be viewed from two perspectives. The estimation form focuses on estimating the value(s) of the function one at a time. That is, the estimator provides a value separately at each data point. Collectively these values determine the interpolated function. Alternatively when the estimator is written in the form given in (15) then one obtains an interpolating function. If all of the data is used then the two approaches are equivalent. Note that the smoothness or continuity of the interpolating function is determined by the smoothness or continuity of the variogram as well as of the monomials.

The number of data locations may vary by several orders of magnitude from one application to another and the coefficient matrix in the linear system is directly related to the number of data locations, hence the matrix may be very large. When the estimator form is used several useful characteristics are noted. First, the coefficients for data values at locations far away are zero or nearly so. Secondly, there is a de-clustering effect. Hence most software packages allow the use of a search neighborhood, i.e., allow limiting the number of data locations used in the estimation. In practice this is less than 30, hence the matrix is much smaller. The trade-off is that a new matrix must be inverted, i.e., a linear system must be solved, for each location where an estimate is desired. This can result in local discontinuities.

There are a few commercial packages that include "kriging" options in an automatic form, i.e, the user is not required to estimate and model the variogram. In general these packages only allow a linear variogram (special case of the power model). In this case the coefficient in the model does not affect the weights in the estimator and hence does not affect the estimated values. It does however affect the kriging variance, these packages generally do not compute the kriging variance. The user may be mis-led when using these since they have not used "kriging" but rather only a special form of it.

C. Symmetry and Other Thoughts

Suppose that all the data locations are on a regular square grid and the objective is to estimate the value at the center of each square using only the data at the four corners of the respective squares. It is easy to see from the kriging equations that if the variogram is isotropic then the weights will all be 0.25. It is also easy to see that most interpolation methods will produce the same results. In this case the same location pattern is more important than the interpolation method and the interpolation method parameters. Symmetry can sometimes be used to simplify the application of the interpolation technique but it can also obscure differences in methods.

Often the data has been composited, i.e., the samples have been composited, or compositing is considered in designing a sampling plan. Compositing may be desirable for a number of reasons, one of which is to reduce the cost of analyzing samples and a second reason is that the analytical method requires a minimum physical sample size but only small physical sample sizes are possible in the sampling process. Compositing will affect the geostatistical analysis in two different ways, one is in the estimation and modeling of variograms and the second is in the application of the (kriging) estimator. These might be viewed as discrete versions of the problem of regularized variograms and the estimation of spatial averages but there are important differences. The potential effects of compositing should be seriously considered before using composited samples in a geostatistical analysis.

The Ordinary and Universal Kriging equations were given in terms of variograms, each can be re-written in terms of the covariance provided a model with a sill is used. This has certain computational advantages when implementing the algorithms in a computer program. Hence most software will internally make this transformation. If a search neighborhood is used then it is possible to make a comparable transformation even if a Power model is used

D. A Few Examples

There have been a large number of examples from a variety of areas of application appearing in the literature in the last 15 years. Myers (1988d) considers the case of hydrogeochemical data. This data was collected as part of the NURE (National Uranium Resource Evaluation) project in the late '70s. The objective had been to search for patterns that would indicate the presence of undiscovered uranium deposits. A companion study using Inverse Distance Weighting appears in Kane et al (1982).

The data used in Myers (1991c) was collected as part of a study of lead concentrations in the soil in Dallas resulting from emissions of a lead smelter. This analysis illustrates the use of the GeoEas package.

Ali et al (1990) used indicator transforms to contour the likelihood of soil collapse in the Tucson valley. This is a serious problem for home construction and is a consequence of the soil formation processes in a semi-arid area.

Zhang et al (1992b, 1992b) show the use of pseudo-cross variograms to improve the interpolation of soil spectral data.

Tabios and Salas (1985) provide empirical comparisons for different interpolation methods for precipitation data.

REFERENCES

Ali,M.,Nowatzki, E. and Myers, D.E., (1991) Probabilistic Analysis of Collapsing soil by Indicator Kriging Math. Geology, 22, 15-38

Cressie, N. 1986, Kriging Non-Stationary Data. J. American Statistical Assoc. 81, 625-634

Cressie, N., 1990, Reply to a letter by G. Wahba. The American Statistician

Davis, B.M. and Borgman, L.E., 1979, Some Exact Sampling Distributions for Variogram Estimators. Math. Geology 11, 643-653

Davis, B.M. and Borgman, L.E., 1982, A Note on the Asymptotic Distribution for Variogram Estimators. Math. Geology 14, 189-194

Dolph, C.L. and Woodbury, M.A. (1952) On the relation between Green's functions and the covariances of certain stochastic processes and its application to unbiased linear prediction. Trans. AMS 519-550

Dowd, P, 1992, A Review of Recent Developments in Geostatistics. Computers and Geostatistics 17, 1481-1500

Dubrule, Olivier (1984) Comparing Splines and Kriging. Computers and Geosciences 10, 327-338

Foley, T.A., (1987) Interpolation and Approximation of 3-D and 4-D Scattered Data. Comp. Math. Applic. 13, 711-740

Goldberger,A.S.,1962, Best Linear Unbiased Prediction in the Generalized Linear Regression Model. J. Am.Stat.Assn., 369-375

Hardy,R.L.,1971, Multiquadratic equations of Topography and other Irregular Surfaces. J. Geophysical Research,76, 1905-1915

Journel, A. G.,1977, Kriging in terms of Projections. Math. Geology 9, 563-586

Journel, A.G. and Rossi, M. 1989, When do we need a Trend model in Kriging. Math. Geology 21, 715-740

Kane,V. C. Begovich, T. Butz and D.E. Myers, (1982) Interpretation of Regional Geochemistry.

Computers and Geosciences, 8, no. 2, 117-136

Kimeldorf,G.S. and Wahba,G.,1970, A Correspondence between Bayesian estimation on Stochastic Processes and Smoothing by Splines. Annals Math. Stat.,41, No. 2, 495-502

Marcotte, D. and David, M., (1988) Trend Surface Analysis as a specical case of IRF-k kriging. Math. Geology 20, 821-824

Matem,B., 1986, Spatial Variation, 2nd Edition, Lecture Notes in Statistics #36, Springer Verlag

Matheron,G.,1965, Les Variables Regionalisees et Leur Estimation. Masson et Cie, Paris

Matheron,G.,1973,The Intrinsic Random Functions and Their Applications. Adv. Applied Prob., 5, 439-468

Matheron,G.,1980, Splines et Krigeage:Leur Equivalence Formelle. Internal Report N-667, Centre de Geostatistique, Fontainebleau, 26p

Matheron,G.,1981a,Remarques sur le Krigeage et son dual. Internal report N-695, Centre de Geostatistique, Fontainebleau, 36p

Matheron,G.,1981b,Splines et Krigeage: Le Cas Fini. Internal Report N-698, Centre de Geostatistique, Fontainebleau,23p

Micchelli,C.,1986,Interpolation of Scattered Data:Distance Matrices and Conditionally Positive Definite Functions. Constructive Approximation, 2, 11-22

Miesch, A.T., 1975, Variograms and Variance Components in Geochemistry and Ore Evaluation. Memoir 142, Geological Society of America, 333-340

Myers,D.E., 1982, Matrix Formulation of Cokriging. Math. Geology, 14, no.3, 249-257

Myers,D.E., 1983, Estimation of Linear Combinations and Cokriging. Math. Geology, 15, no.5, 633-637

Myers,D.E., 1984, Cokriging:New Developments. in Geostatistics for Natural Resource Characterization, G.Verly et al, eds., D. Reidel Pub. Co., Dordrecht, 295-305

Myers,D.E., 1988a,Interpolation with Positive Definite Functions. Sciences de la Terre, 28, 251-265

Myers,D.E., 1988b, Multivariate Geostatistics for Environmental Monitoring. Sciences de la Terre, 27, 411-427

Myers,D.E., 1988c, Some Aspects of Multivariate Analysis. in Quantitative Analysis of Mineral and Energy Resources, C.F. Chung et al (eds), D. Reidel Publishing Co., Dordrecht, 669-687

Myers, D.E., 1988d, Kriging Hydrogeochemical Data, in Current Trends in Geomathematics, D.F. Merriam (ed) Plenum Press 117-142

Myers,D.E., 1989a, To Be or Not to Be...Stationary:That is the Question. Math. Geology, 21, 347-362

Myers, D.E. ,1989b, Borden Field Data and Multivariate Geostatistics. in Hydraulic Engineering, M.A. Ports (ed), Amer. Soc. Civil Eng. 795-800

Myers, D.E., 1991a On Variogram Estimation. in Proceedings of the First Inter. Conf. Stat. Comp., Cesme, Turkey, 30 Mar.-2 April 1987, Vol II, American Sciences Press, 261-281

Myers,D.E., 1991b Multivariate, Multidimensional Smoothing. in Spatial Statistics and Imaging: Proceedings of an AMS-IMS-SIAM Joint Summer Research Conference, June 18-24, 1988, Bowden College, Maine. Lecture Notes-Mongraph Series, Institute of Mathematical Statistics, Hayward, CA 275-285

Myers,D.E., 1991c Interpolation and Estimation with Spatially Located Data, Chemometrics and Intelligent Laboratory Systems, 11, 209-228

Myers,D.E., 1991d Pseudo-Cross Variograms, Positive Definiteness and Cokriging. Math. Geology 23, 805-816

Myers, D.E.1992, Kriging, Cokriging and the Role of Positive Definiteness. Computers Math. Applications 24, 139-148

Myers,D.E.,1993 , Selection of a Radial Basis Function for Interpolation. in Advances in Computer Mathods for Partial Differential Equations-VII, R. Vichnevetsky, D. Knight and G. Richter, International Association for Mathematics and Computers in Simulation, New Brunswick 553-558

Norris, M.D., 1991, On counting the number of Data Pairs for Semivariogram Estimation. Math. Geology 23, 929-944

Parker, H. 1979, The Volume-Variance Relationship:A Useful Tool for Mine Planning. Eng. & Mining Journal, October, 106-123

Powell,M.J.D.,1985, Radial Basis Functions for Multivariable Interpolation:A Review. in Algorithms for the Approximation of Functions and Data, M.G. Cox and J.C.Mason, (eds), Oxford University Press, Tabios, G.Q III and Salas, J.D., 1985, A Comparative Analysis of Techniques for Spatial Interpolation of Preciptitation. Water Resources Bulletin 21, 365-380

Wahba,G.,1975, Smoothing Noisy Data with Spline Functions. Numer. Math., 24, 383-393

Warrick A. and D.E. Myers, (1987) Optimization of Sampling Locations for Variogram Calculations. Water Resources Research, 23, 496-500

Watson,G.S.,1984, Smoothing and Interpolation and Splines. Math. Geology, 16,No.8, 601-615

Zhang, R., Warrick, A.W. and Myers, D.E., (1990) Variance as a Function of Sample Support Size. Math. Geology 22, 107-121

Zhang,R. D.E. Myers and A.W. Warrick, 1992a, Estimation of the Spatial Distribution of Soil Chemicals using Pseudo-Cross Variograms. Soil Science Society of America Journal 56, 1444-1452

Zhang, R. D.E. Myers and A. Warrick,1992b, Improvement of the prediction of particle size fractions using spectral properties. Geoderma 22, 223-234

Zimmerman, D.L. and Zimmerman, B. 1991, A Comparison of Semivariogram Estimators and Corresponding Ordinary Kriging Predictors. Technometrics 33, 77- 91