

ESTIMATION OF CORRELATION STRUCTURE FOR A HOMOGENEOUS ISOTROPIC RANDOM FIELD: A SIMULATION STUDY

WITOLD F. KRAJEWSKI¹ and CHRISTOPHER J. DUFFY²

¹Iowa Institute of Hydraulic Research, The University of Iowa, Iowa City, IA 52242 and ²Civil Environmental Engineering Department, Utah Water Research Laboratory, Utah State University, Logan, UT 84322, U.S.A.

(Received 1 May 1987; revised 10 August 1987)

Abstract—In this paper we investigate the effect of sampling density on the estimation of the covariance and semivariogram for homogeneous, isotropic, random fields. Two methods based on the least-squares principle, and a third method known as the Minimum Interpolation Error method are studied when the analytic form of the covariance or semivariogram model is known a priori. The analysis is accomplished through a single realization simulation experiment which is felt to represent the type of conditions usually encountered in real world environmental and geophysical field problems. The Turning Bands method is used to generate the field at randomly distributed sampling points in a fixed field for three types of correlation structure: exponential, Bessel, and Gaussian models. The performance of the three estimation methods is evaluated for varying sampling densities and correlation distances. The main results are: the least-squares methods work best for preserving the pattern of correlation in most situations examined; for a domain of fixed size, the ratio of the correlation distance to the length scale of the field is a measure of the "information" contained in the field, and when this ratio exceeded ≈ 0.2 the statistics of the process became inaccurate. On the other hand, when this ratio is ≤ 0.2 reasonable estimates for the mean and variance were determined even for small sampling densities ($\approx 25-50$). The implications for practical problems are discussed.

Key Words: Random fields, Covariance estimation, Sensitivity.

INTRODUCTION

The analysis of geophysical phenomena as random field now has become a widely used method for characterizing spatial data. In groundwater hydrology the research of Gambolati and Volpi (1979) and Kitanidis (1983) are examples of spatial characterization and interpolation of hydraulic properties in porous media. The work of Chua and Bras (1982) and Creutin and Obled (1982) provide other examples applied to the study of rainfall fields. Another area where characterization of field scale variability has emerged as a critical problem is transport of contaminants or tracers in groundwater systems. Theories of multidimensional dispersive mixing have been proposed recently that suggest the variance and correlation structure of permeability is an essential element of the dispersion process. The work of Matheron and deMarsily (1980), Gelhar and Axness (1983), and Neuman, Winter, and Neuman (1987) stand out in this area. This research has stimulated a number of field experiments (e.g. MacKay and others, 1986), to test the validity of the proposed theories. In each situation the ability to estimate accurately the correlation structure and statistical properties of the porous medium from sampled field data is required.

For many applications the problem of sparse data limits our ability to make meaningful estimates of the statistics of the field or its correlation structure. Re-

sults from time-series analysis offer some guidance to this question, for example Jenkins and Watts (1968, p. 53) state, "... the length T of the record determines the extent to which peaks in the Fourier transform (of the autocovariance function) may be distinguished. On the other hand, the sampling interval Δ determines the maximum frequency (Nyquist frequency) which can be detected." Thus, it has been determined that the record length (or size of the domain), and the average sampling interval (or sampling density) are interrelated; and given some information about the system at hand, an experiment actually can be designed for a desired resolution and accuracy by controlling T and Δ .

For spatially random processes these same considerations certainly are valid, however, unlike time-series analysis we have less control over the experiment, because the domain may be fixed by nature (e.g. a soil type or geologic strata), and the sampling density limited by economic considerations.

In this paper we examine the question of estimation of correlation structure for two-dimensional, stationary, random fields, as a function of sampling density for a domain of fixed area. Our approach has been to conduct a fully controlled simulation experiment to determine the behavior of estimators for the mean, variance, covariance, and semivariogram under changing sampling density and correlation length. In this situation the sampling scheme is taken

to be uniformly random. Although we recognize that under certain conditions there can be more efficient sampling schemes (Ripley, 1982), the assumption is justified here based on practical considerations such as: (a) many existing networks can be described as approximately uniformly random distributed; and (b) as has been pointed out by Masry (1971), the reconstruction of spectral and covariance estimators for time series which minimize aliasing, is accomplished more efficiently by random sampling schemes. The drawback to random sampling schemes, of course, is the computational difficulty of estimating second-order structure functions from scattered data. However, because scattered data is a common feature of field problems, this will be the main concern of the paper.

In the experiment to follow we limit our considerations to stationary or more correctly, homogeneous random fields, in two-dimensional space with isotropic correlation structure. The experiment is conducted by generating a single realization of the random field for a specified number of sampling points via the Turning Bands method (Mantoglou and Wilson, 1982). From each realization the parameters of a specified correlation structure are estimated using one of three methods. The first two estimation methods are variations of the least-squares technique, and the third method is based on the Minimum Interpolation Error method (Bastin and Gevers, 1985). In each situation the form of the theoretical covariance or semivariogram is assumed to be known a priori. Extension of the analysis to model identification is left for future work.

Because the functional form of the true covariance of the underlying process and its parameter values are known, we can compare the performance of our estimators with respect to this true covariance. The problem of inferring the covariance of the process from a single realization is of practical interest if the estimated covariance is to be used in simulation-type studies. However, the estimators presented here also can be used to estimate the realization-specific covariance.

BASIC DEFINITIONS

A random process $Z(\mathbf{u})$ is defined over the domain $\Omega \subset R^2$ with $z(\mathbf{u})$ representing a single realization of the process and $\mathbf{u} = (x, y)$ denotes location in Ω . The mean of the process $Z(\mathbf{u})$ is given by

$$\mu(\mathbf{u}) = E\{Z(\mathbf{u})\} \quad \mathbf{u} \in \Omega \quad (1)$$

where $E\{\cdot\}$ is the expectation operator. The variance of $Z(\mathbf{u})$ is

$$\sigma^2(\mathbf{u}) = \text{Var}\{Z(\mathbf{u})\} = E\{[Z(\mathbf{u}) - \mu(\mathbf{u})]^2\} \quad (2)$$

and the covariance between two points $\mathbf{u}_1 = (x_1, y_1)$ and $\mathbf{u}_2 = (x_2, y_2)$ is defined as

$$\text{Cov}(\mathbf{u}_1, \mathbf{u}_2) = E\{[Z(\mathbf{u}_1) - \mu(\mathbf{u}_1)][Z(\mathbf{u}_2) - \mu(\mathbf{u}_2)]\}. \quad (3)$$

From the geostatistics literature the semivariogram is defined as

$$\gamma(\mathbf{u}_1, \mathbf{u}_2) = \frac{1}{2} \text{Var}\{Z(\mathbf{u}_1) - Z(\mathbf{u}_2)\}. \quad (4)$$

The field $Z(\mathbf{u})$ is said to be homogeneous statistically (or second-order stationary) and isotropic if the mean is constant and

$$\text{Cov}(\mathbf{u}_1, \mathbf{u}_2) = \text{Cov}(\mathbf{u}_1 - \mathbf{u}_2) = \text{Cov}(\mathbf{r}) = c(r) \quad (5)$$

where

$$\mathbf{r} = \mathbf{u}_1 - \mathbf{u}_2.$$

This implies that the covariance function depends only on the distance between the points and not on direction in the field.

Journel and Huijbregts (1978) and Mantoglou and Wilson (1982) give several forms of covariance functions used in geophysical analysis. The following are adopted in this study:

(1) Exponential

$$c(r) = \sigma^2 \exp(-Br) \quad r > 0 \text{ and } B \geq 0. \quad (6)$$

(2) Bessel-type

$$c(r) = \sigma^2 Br K_1(Br) \quad r \geq 0 \quad (7)$$

where $K_1(\cdot)$ is modified Bessel function of second type of order 1.

(3) Gaussian-type

$$c(r) = \sigma^2 \exp(-B^2 r^2) \quad r \geq 0. \quad (8)$$

In each situation the parameter B is proportional to the inverse of the so-called "correlation scale" of the process, which in practical terms is a measure of the distance over which the process is correlated. We will refer to B as the correlation parameter and B^{-1} as the correlation length. Figure 1 illustrates the covariance models used in the study.

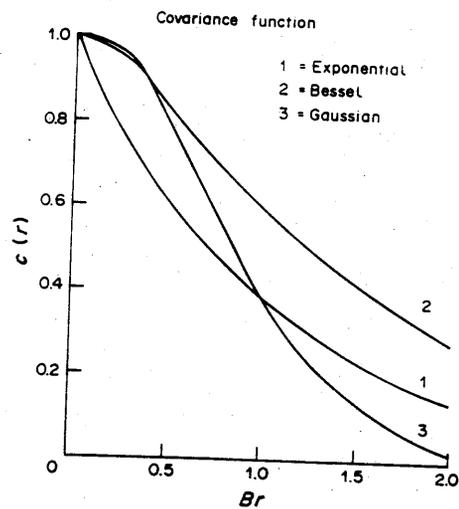


Figure 1. Covariance functions used in simulation study.

It can be demonstrated (Journel and Huijbregts, 1978) that for homogeneous fields the semivariogram and the covariance are related

$$\gamma(r) = c(0) - c(r) \quad (9)$$

or in terms of the correlation function $\varrho(r)$

$$\varrho(r) = 1 - [\gamma(r)/c(0)]. \quad (10)$$

It follows from Equations (9) and (10) that for a statistically homogeneous field, the semivariogram γ corresponding to Equations (6)–(8) is given by

$$\gamma_1(r) = \sigma^2[1 - \exp(-Br)] \quad B > 0 \quad r \geq 0 \quad (11)$$

$$\gamma_2(r) = \sigma^2[1 - BrK_1(Br)] \quad r \geq 0 \quad (12)$$

$$\gamma_3(r) = \sigma^2[1 - \exp(-B^2r^2)] \quad r \geq 0. \quad (13)$$

The main objective of this study then is to estimate parameters of the covariance or semivariogram from a single realization of the field, for variable sampling densities.

ESTIMATION OF THE FIELD STRUCTURE

In reality a random field is sampled at a discrete number of locations in space. Ideally, one would prefer to have a number of realizations of this field, however, in most instances this is not possible. Thus, to be able to estimate the parameters of the field and taking advantage of the underlying theory, we make the operational assumption that the statistics of a single realization are approximately the same as those of the underlying random field.

At this point we construct the estimators of the field sampled at n locations. Let $U = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ be a set of n sampling points in R^2 with $z(\mathbf{u}_i)$ for $i = 1, \dots, n$ being the corresponding values of the realization. These points are taken from a random, uniform distribution in the region $\Omega \subset R^2$. It is assumed throughout the paper that the observations $z(\mathbf{u}_i)$ are error free. The mean μ of random field Z now can be estimated as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i) \quad (14)$$

and the variance σ^2 as

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n \{z(\mathbf{u}_i) - \hat{\mu}\}^2. \quad (15)$$

In this paper, estimation methods of field covariance or semivariogram functions are based on two different approaches. The first approach requires construction of the empirical covariance (or semivariogram) which then is approximated by a particular model. The second approach is direct in that no empirical statistics of the covariance are required and the parameters of the assumed model are estimated directly from the sample points.

In the first approach we construct the empirical covariance (or semivariogram) by looking for all pairs of data points $(\mathbf{u}_i, \mathbf{u}_j)$ separated by distance $r_k \pm \Delta$, for $k = 1, \dots, K$, where K is the number of intervals for which we want to compute the empirical correlation structure. Then, for all the pairs we form the empirical covariance:

$$c_k(r_k) = \frac{1}{n_k} \sum_i \sum_j [z(\mathbf{u}_i) - \hat{\mu}][z(\mathbf{u}_j) - \hat{\mu}] \quad \forall_{i,j} \{i, j: \mathbf{u}_i - \mathbf{u}_j \in \langle r_k - \Delta, r_k + \Delta \rangle\} \quad (16)$$

where n_k is the number of pairs satisfying the condition given in Equation (16). Similarly the empirical semivariogram is:

$$\gamma_k(r_k) = \frac{1}{n_k} \sum_i \sum_j [z(\mathbf{u}_i) - z(\mathbf{u}_j)]^2 \quad \forall_{i,j} \{i, j: \mathbf{u}_i - \mathbf{u}_j \in \langle r_k - \Delta, r_k + \Delta \rangle\}. \quad (17)$$

As we can see from Equation (17), computation of the variogram does not require the knowledge of the mean. The consequences of this fact will be discussed subsequently.

The parameters of a particular covariance (or semivariogram) function given by Equations (6), (7), or (8) can be obtained by minimization of

$$\sum_{k=1}^K [c(r_k) - c_k(r_k)]^2 \quad (18)$$

with respect to the vector of parameters $\mathbf{a} = \langle B \rangle$. Similarly for the variogram, minimization of

$$\sum_{k=1}^K [\gamma(r_k) - \gamma_k(r_k)]^2 \quad (19)$$

yields estimates of parameters for the semivariogram models as given in Equations (11), (12), and (13).

The basic question that arises from implementation of this scheme is how to select r_k and Δ . The interval $(r - \Delta, r + \Delta)$ should be small enough to minimize the smoothing effect introduced by averaging overall pairs included in it, and large enough so that sufficient number of pairs n_k are available to obtain reliable estimates of $c_k(r_k)$ or $\gamma_k(r_k)$. This can be a critical problem particularly at small lag distances if the covariance (or semivariogram) is to be used for interpolation of the field. Unfortunately, because the average number of pairs in each interval is proportional to the distance, there may be a problem obtaining enough data at small distances. Journel and Huijbregts (1978) suggest that at least 30–50 pairs should be used.

In this study we investigated two different approaches for settling r and Δ .

Scheme 1. It is assumed that:

- (1) $r_{k+1} - r_k = \text{constant}$ $k = 1, \dots, K-1$.
 (2) $r_1 = \frac{1}{n} \sum_{i=1}^n \min(d_{ij}; j = 1, \dots, n; j \neq i)$

where d_{ij} is the distance from point u_i to point u_j .

- (3) $\Delta = r_1/2$.
 (4) K is fixed (usually $K = 10-15$).

In other words r_1 is an average of the minimum distances between the points in the field, and for this example all the intervals are equal. In this scheme we do not have any control of the number of pairs n_k for k -th interval.

Scheme 2. This method is based on the following assumptions:

- (1) $r_k = \min\{r_k: n_k(r_k) \geq N^*\}$ for $k = 1, \dots, K$.
 (2) $r_{k+1} - r_k \geq r_k - r_{k-1}$ for $k = 2, \dots, K-1$.
 (3) $\Delta_k = r_k/2$ for $k = 1, \dots, K$.
 (4) $K = \min \left\{ K_{\max}, \min \left\{ k: \sum_{i=1}^k r_i \geq D^* \right. \right. \\ \left. \left. \text{for } k = 1, \dots, K_{\max} \right\} \right\}$

where D^* is one-half of the maximum distance in Ω , K_{\max} is fixed ($K_{\max} = 10-15$), and N^* is the minimum required number of pairs, the same for each interval.

We can describe this method in the following way. Each interval length is minimized with the restriction that it includes a minimum number of pairs (not less than N^*). Subsequent intervals are at least as large as each previous one, and the total number of intervals is such that it covers no more than one-half of the maximum distance in Ω .

Once the empirical covariances (or variograms) are computed we need to fit a selected model as in Equations (6), (7), or (8) using Equations (18) or (19). As was mentioned earlier, we will use the correct model (i.e. the one from which the field was generated) to determine the effect of sampling density on the estimation of the correlation in the field.

The second approach to estimate the field covariance (or semivariogram) is a direct method which does not require computation of empirical functions. The method is known as the Minimum Interpolation Error method and is described by Ripley (1981), Lebel and Bastin (1985), and extensively discussed by Bastin and Gevers (1985). It is a cross-validation-type method where each data point is withheld in turn and its value is interpolated from the neighboring point. Formally, it can be stated as:

$$\min_{\mathbf{a}} \sum \{z(u_i) - \hat{z}(u_i, \mathbf{a})\}^2 \quad (20)$$

where $\hat{z}(u_i, \mathbf{a})$ is an interpolated value of field Z at the point u_i . It also is a function of parameters \mathbf{a} (in our situation a single parameter B) through the covariance (or semivariogram) model and can be expressed as:

$$\hat{z}(u_i, \mathbf{a}) = \sum_{j=1}^J \alpha_j z(u_j) \quad (21)$$

for $i = 1, \dots, n$ and $j \neq i$.

The weights α_j in the linear estimator of Equation (21) can be determined by minimizing the variance of \hat{z} (e.g. see Journel and Huijbregts, 1978) under nonbias condition:

$$\sum_{j=1}^J \alpha_j = 1. \quad (22)$$

This leads to the system:

$$\begin{vmatrix} \text{Cov}(u_1, u_1) & \dots & \text{Cov}(u_1, u_j) & 1 \\ \vdots & & \vdots & \vdots \\ \text{Cov}(u_j, u_1) & \dots & \text{Cov}(u_j, u_j) & 1 \\ 1 & \dots & 1 & 0 \end{vmatrix} \begin{matrix} x_1 \\ \vdots \\ x_j \\ \mu \end{matrix} = \begin{vmatrix} \text{Cov}(u_1, u_i) \\ \vdots \\ \text{Cov}(u_j, u_i) \\ 1 \end{vmatrix} \quad (23)$$

where μ is a Lagrange multiplier, and $i = 1, \dots, n$.

The number of neighboring points, J , is selected arbitrarily, however, the points should be those with the highest correlation with the estimated point, that is in the situation of isotropic fields, those closest. In our study, J was fixed and equal to 4. Later we will discuss some of the consequences of this assumption in more detail. At this point note that the direct method described here becomes expensive computationally if there are many data points. Also, the criterion of Equation (20) is not objective because it may be dominated by a few data points that are difficult to predict because of their isolation (Ripley, 1981).

GENERATION OF THE RANDOM FIELD

In generating a random field it is first necessary to select a sampling scheme. In the time-series context, Masry (1971) has pointed out that regular uniform sampling may be inadequate for a unique reconstruction of the spectrum or the covariance structure from sampled data. Rather, he demonstrates that random sampling is a more fruitful approach, and goes on to suggest a number of possible schemes. It seems reasonable that random sampling is appropriate for multidimensional fields as well, and without the be-

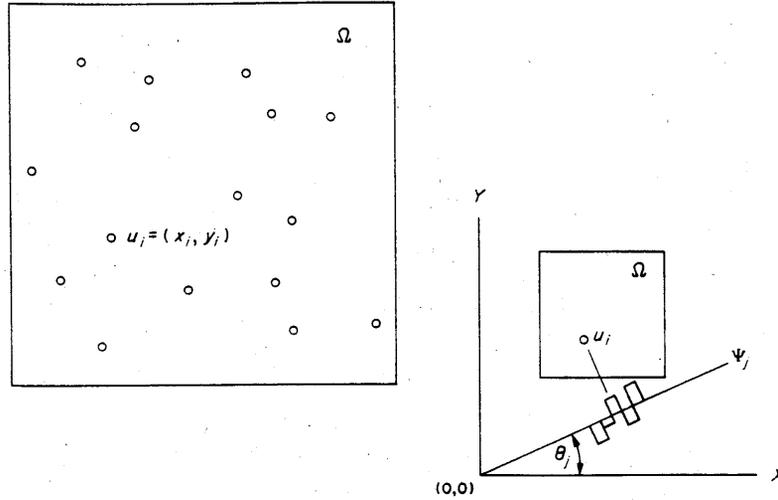


Figure 2. Generation of 2-D random field from 1-D process along lines Ψ_j .

nefit of a rigorous justification we adopt one such approach here. We use a random sampling scheme where the coordinates x_i and y_i of points $u_i = (x_i, y_i)$, for $i = 1, 2, \dots, n$ are drawn from the uniform distributions $U(X, 0)$ and $U(0, Y)$ over the rectangular region Ω with area $X \cdot Y$. The second step is to generate values of the random process $z(u_i)$ at each location in the field, $u_i, i = 1, 2, \dots, n$. This is done using the Turning Bands Method (TBM), a name coined by Matheron (1973). Mantoglou and Wilson (1982) derive the TBM equations for the two-dimensional situation, which is recounted briefly here. The basic idea is to transform a stationary, two-dimensional random field into the sum of a series of equivalent one-dimensional line processes. According to Mantoglou and Wilson (1982), independent realizations of the line process Ψ_j with mean μ and covariance $c_1(\tau)$, are generated along L intersecting lines (see Fig. 2). For each location $u_i, z(u_i)$ is estimated from

$$z(u_i) = \frac{1}{\sqrt{L}} \sum_{j=1}^L \Psi_j \quad (24)$$

where L is the total number of intersecting lines. Mantoglou and Wilson provide details of the transformation between the covariance of the line process $c_1(\tau)$ and the general two-dimensional field.

The actual generation of the line process is accomplished here using the spectral method of Shinozuka and Jan (1972) from the expression

$$\Psi_j(\tau) = 2 \sum_{k=1}^m [s_1(\omega_k) \Delta\omega]^{1/2} \cos(\omega_k \tau + \Phi_k) \quad (25)$$

where $s_1(\omega_k)$ is the spectral density for the covariance $c_1(\tau)$. The spectrum is discretized into m components of central frequencies $\omega_k, k = 1, \dots, m$ with the increments $\omega_k - \omega_{k-1} = \Delta\omega$ independent of k . Φ_k are independent random phase angles, uniformly dis-

tributed on the interval $\langle 0, 2\pi \rangle$. The frequency ω_k is the sum of ω_k and a small random frequency $\delta\omega$, uniformly distributed on $\langle -\Delta\omega/2, +\Delta\omega/2 \rangle$ with $\Delta\omega_k \ll \Delta\omega$. The addition of this small random frequency is to avoid introduction of periodicities in the result.

The spectral densities which correspond to Equations (6), (7), and (8) along an arbitrary line in the field are given by

(1) Exponential

$$s_1(\omega) = \frac{\sigma^2}{2} \frac{\omega}{B^2(1 + \omega^2/B^2)^{3/2}} \quad (26)$$

(2) Bessel-type

$$s_1(\omega) = \sigma^2 \frac{\omega}{B^2(1 + \omega^2/B^2)^2} \quad (27)$$

(3) Gaussian-type

$$s_1(\omega) = \frac{\sigma^2}{4B^2} \exp(-\omega^2/4B^2) \quad (28)$$

The expressions along with Equations (22) and (24) are used to generate the field for a specified sampling density.

SIMULATION RESULTS

Performance criteria

Because we assume that the covariance function is known, the objective is to estimate the parameters of each generated field, namely the mean, variance, and correlation scale. Then we determine the performance of each estimator relative to the original process as a function of sampling density. The performance of the mean and variance estimators is straightforward because the original process has $\mu = 0$ and $\sigma^2 = 1$. Although the parameter B for each process is proportional to the inverse of correlation length B^{-1} [length],

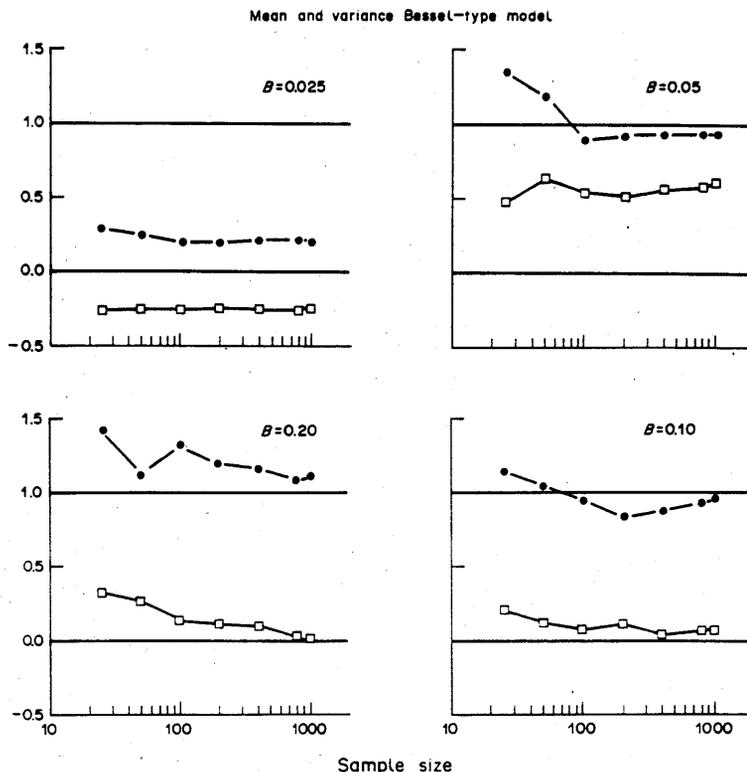


Figure 3. Estimates of mean and variance vs sampling density for Bessel correlation model. ●—Variance estimate; □—mean estimate.

it is convenient in this example to use $(\hat{B} - B)$ as our performance measure, instead of $\hat{B}^{-1} - B^{-1}$, where \hat{B} is the estimate and B is the true value.

Some practical results

For this experiment we generated realizations of $Z(u_i)$, for $n = 25, 50, 100, 400, 800$, and 1000 points in a field Ω of 100×100 units. The shape parameter B was specified as 0.025, 0.05, 0.10, and 0.20, which corresponds to length scales of 40, 20, 10, and 5 units. Although it is not necessary for the Turning Bands Method, fixing the size of the field at 100×100 units allows us to examine the effect of domain size on the estimators. For many practical problems the observed field itself may be fixed arbitrarily, either from physiographic or economic considerations (e.g. agricultural field soils, physical boundaries, or landforms which are too extensive and thus too expensive for detailed field investigations, or because the domain is simply unknown prior to sampling).

Figure 3 illustrates the effect of sampling density on the mean and the variance estimates for realizations generated from the Bessel-type correlation model as shown in Equation (12). The mean and variance estimators for the Gaussian and exponential models show similar behavior. For the situation of large B (0.10, 0.20) or small correlation lengths B^{-1} (20, 5) both the mean and variance estimates converge to the expected values as n gets large. However, for

small values of B (0.025, 0.05) or large correlation lengths (40, 20) the performance of the estimators is poor with larger errors in the mean or variance and no clear convergence as n gets large. Similar results were observed for the exponential and Gaussian models. Although we determined no differences in estimating the mean and variance between the random field models under consideration, it became clear that the critical factor in parameter estimation was the degree of correlation in the field. In the situation of short-range correlation (large B), the first two moment estimators performed well, whereas for long-range correlation (small B) the estimators demonstrated a poorer performance. An explanation of the poor performance of the estimator for long-range correlation may be determined from sampling theory in time-series analysis, where it is well known that a precise reconstruction of the covariance or spectrum and thus the variance of the process, is not possible when the decorrelation time (B^{-1} in our situation) approaches the length of the record (length of record $L \approx \sqrt{\text{area}}$ in our example). The problem is that by fixing the size of the sampled field for a process with continuous covariance, information from frequencies in the record $\omega > 4\pi/L$ are lost in the estimation. How large an impact this has depends on the correlation scale of the process. Figure 4 illustrates this result for sampling densities of $n = 100$ and 1000 for each of the correlation models used. As long as the length scale

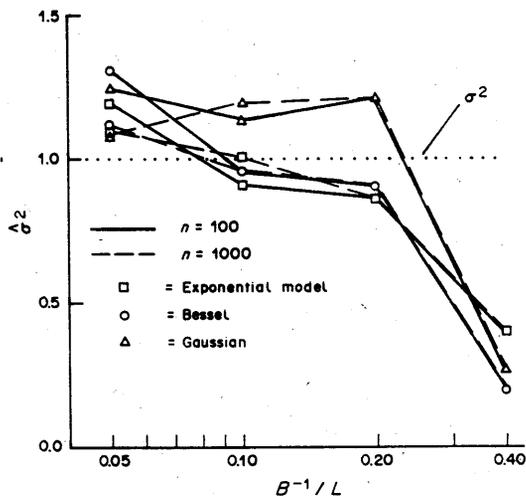


Figure 4. Variance estimate vs correlation length scale.

B^{-1} did not exceed 20% of the length scale of the field ($L = 100$ units), reasonably good estimates of the variance could be determined even for sample sizes of 25–100. However, when B^{-1} was 40% of the field size, even sampling densities of 1000 random points showed poor estimator performance.

Estimating the shape parameter B

The shape parameters for each covariance and semivariogram model [Eqs. (6)–(8) and (11)–(13)] also were estimated as a function of sampling density using the three methods outlined earlier: (a) a least-squares method where the lag interval is fixed and thus the number of pairs per interval is arbitrary, (b) a least-squares method where the lag interval is selected so that there are a minimum number of pairs per interval, and (c) a direct estimation approach known as the Minimum Interpolation Error method. Figures 5, 6, and 7 summarize the results of the numerical experiment.

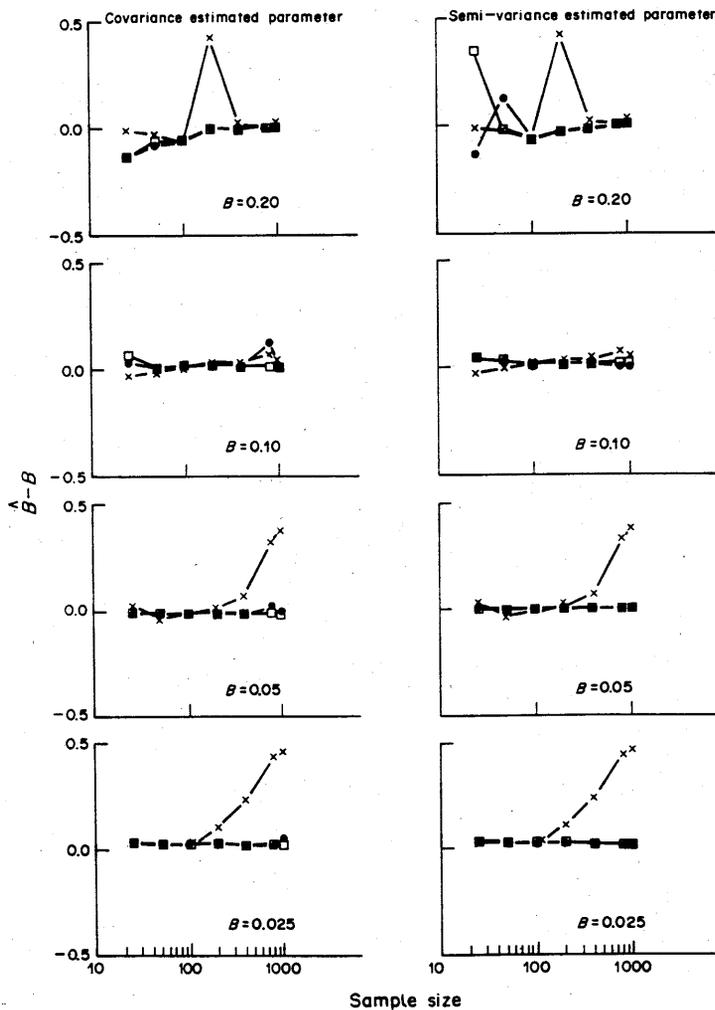


Figure 5. Shape parameter error $\hat{B} - B$ vs sampling density for Gaussian-type correlation model. \square —Least-squares method 1; \bullet —least-squares method 2; \times —Minimum Interpolation Error method.

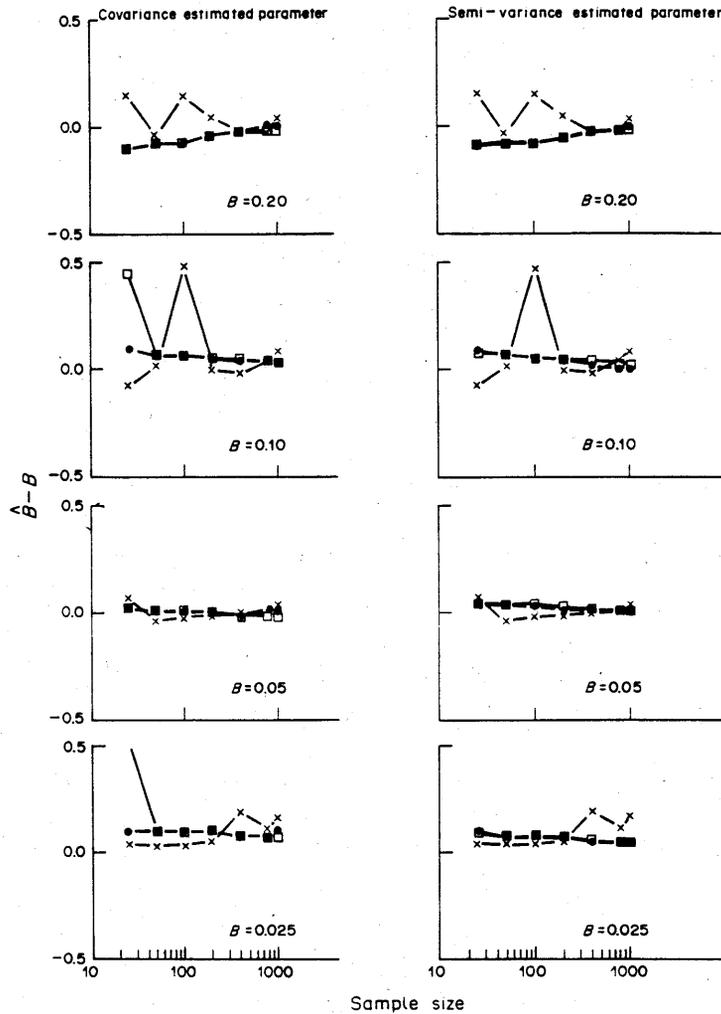


Figure 6. Shape parameter error $\hat{B} - B$ vs sampling density for the Bessel-type correlation model. \square —Least-squares method 1; \bullet —least-squares method 2; \times —Minimum Interpolation Error method.

The first observation we can make is with regard to the estimation of B from the covariance [Eqs. (6)–(8)] or the semivariogram [Eqs. (11)–(13)] models. Because the covariance requires a priori knowledge of the mean we might expect bias in this additional information to affect the estimation of B . However, for our experiments there was almost no difference between the B estimates from the covariance and the semivariogram models, and this result seems to be the situation whether the original field was Gaussian, Bessel, or exponential in structure. This would suggest that for the homogeneous (stationary) fields examined here, there is no practical disadvantage to using either the covariance or semivariogram to estimate correlation structure even though the covariance requires the estimation of an additional parameter. Of course this experiment says nothing about parameter estimation for nonstationary fields, or for weaker stationary assumptions such as stationary increments.

A second observation is that, for the exponential, Gaussian, and Bessel correlation models, the least-

square methods for estimation of B consistently converge ($\hat{B} - B \rightarrow 0$) as sampling density increases. Whether we used a constant lag interval (arbitrary number of pairs) or a variable lag interval (minimum number of pairs) made no discernable difference in estimating B . The Minimum Interpolation Error method (MIE) as applied here produced mixed results. In the Gaussian situation (Fig. 5), the B estimate diverged for several examples ($B = 0.05, 0.025$) as sampling density became large. For all three models MIE had a tendency to oscillate about the true value, whereas the least-square method demonstrated a monotonic convergence with increasing sampling density. Such behavior can be explained here by the way in which we implemented MIE method. We included only four closest predictors [$J = 4$ in Eq. (21)]. For high sampling density the correlation scale of the field is observed through these closest points only. Therefore, the behavior of the correlation models near the origin (Fig. 1) becomes critical with respect to the method's ability to distinguish between various

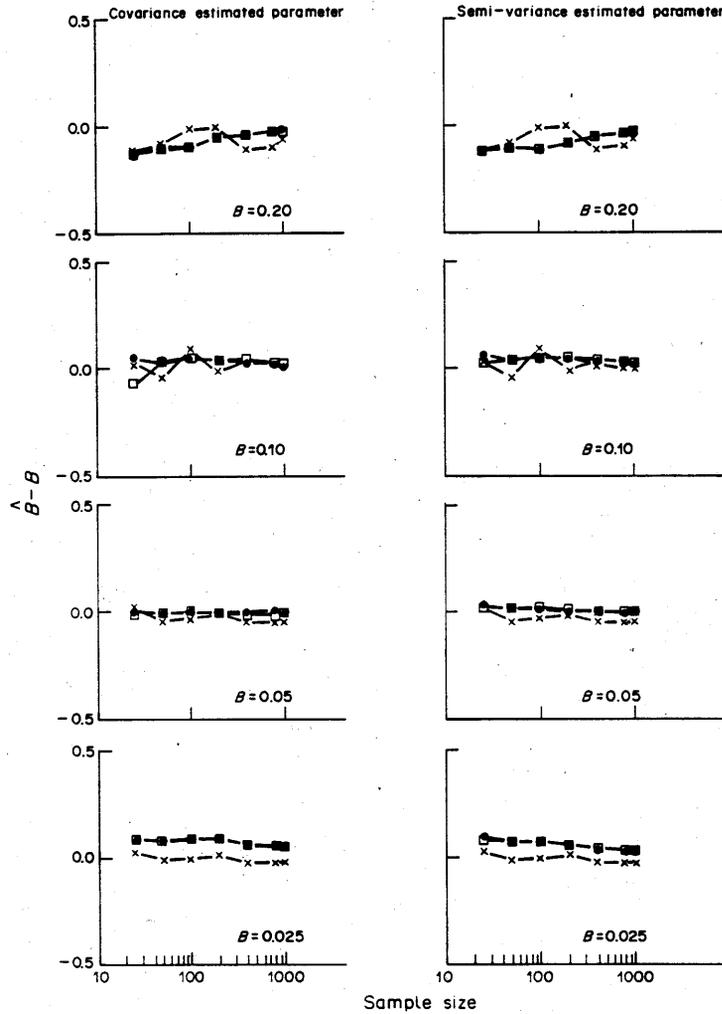


Figure 7. Shape parameter error $\hat{B} - B$ vs sampling density for exponential-type correlation model. \square —Least-squares method 1; \bullet —least-squares method 2; \times —Minimum Interpolation Error method.

correlation scales. This effect is apparent in Figures 5–7 where the model with the smoothest behavior near the origin (Bessel) gives the worst results for higher sampling density. According to Bastin and Gevers (1985) the MIE method can be improved via the maximum likelihood approach, however, a distribution assumption on the errors $\hat{B} - B$ must be made. In general our results suggest that the least-square techniques provide a “better” estimate of the correlation parameter B , or in other words the pattern of correlation is preserved most accurately.

SUMMARY AND CONCLUSIONS

A single realization simulation experiment was conducted to determine the effect of sampling density on the estimation of correlation structure in a homogeneous and isotropic random field. For a sampled field of fixed size the estimated variance is affected apparently strongly by the magnitude of spatial corre-

lation relative to the size of the field itself. When the ratio of correlation length to field size exceeded $B^{-1}/L \approx 0.20$, the estimated variance became inaccurate even for large sample size ($n = 1000$). Conversely, when this ratio was < 0.20 the estimated variance of the field was close to the true value even for relatively small sample size ($n = 25-100$). The effect of long-range correlation on field sampling problems where the domain is fixed arbitrarily in size (e.g. soil survey in agricultural lands, resource reserve estimation, etc.) could be dramatic; because the basic statistics of the overall process would not be estimated accurately regardless of the number of samples. The implication of long-range correlation to Monte Carlo experiments where parameter fields are generated may be more critical. For example there has been interest recently in evaluating the impact of spatial variations in permeability on hydraulic head in aquifers (Smith and Freeze, 1979; and others). In this situation the generated parameter field is used to assess the mean, variance, and correlation structure of the hydraulic

head. Because the filter used is of the "low-pass" type, the output process head, naturally will have a longer correlation length scale than the input process, permeability. If the correlation structure of the input is not selected carefully, the moments and correlation of the output process will not be estimated accurately no matter the number of points in the generated field.

For the homogeneous example, estimation of the correlation parameter B as a function of sampling density indicated no difference in using either the covariance or variogram estimators. Both estimators produced nearly identical results. However we would not expect this to be the situation for nonstationary fields where the mean or trend must be estimated as well.

The study performed here has shown the usefulness of simulation type experiments for investigation of random fields. It has provided some quantitative background for the qualitative problem of structure estimation well known from mathematical and statistical analysis. Similar studies on model identification and nonstationary estimation are planned as a natural extension of this work.

Acknowledgments—The research was supported by the National Science Foundation (CEE-8307982) and the Hydrologic Research Laboratory of the National Weather Service, Silver Spring, Maryland, where the numerical computations were performed. This support is acknowledged gratefully.

REFERENCES

- Bastin, G., and Gevers, M., 1985. Identification and optimal estimation of random fields from scattered point-wise data: *Automatica*, v. 21, no. 5, p. 139–155.
- Chua, S. H., and Bras, R. L., 1982. Optimal estimators of mean area precipitation in regions of orographic influence: *Jour. Hydrology*, v. 57, no. 112, p. 23–48.
- Creutin, J. D., and Obled, C., 1982. Objective analysis and mapping techniques for rainfall fields: an objective comparison: *Water Resources Res.*, v. 18, no. 2, p. 413–431.
- Gambolati, G., and Volpi, G., 1979. Groundwater contour mapping in Venice by stochastic interpolators: *Water Resources Res.*, v. 15, no. 2, p. 281–297.
- Gelhar, L. W., and Axness, C. L., 1983. Three-dimensional stochastic analysis of macrodispersion in aquifers: *Water Resources Res.*, v. 19, no. 1, p. 161–180.
- Jenkins, G. M., and Watts, D. G., 1968. *Spectral analysis and its applications*: Holden-Day, San Francisco, 525 p.
- Journel, A. G., and Huijbregts, C. I., 1978. *Mining geostatistics*: Academic Press, New York, 597 p.
- Kitanidis, P. K., 1983. Statistical estimation of polynomial generalized covariance functions and hydrologic applications: *Water Resources Res.*, v. 19, no. 4, p. 909–921.
- Lebel, T., and Bastin, G., 1985. Variogram identification by the mean-squared interpolation error method with application to hydrologic fields: *Jour. Hydrology*, v. 77, no. 1, p. 31–56.
- MacKay, D. M., Freyberg, D. L., Roberts, P. V., and Cherry, J. A., 1986. A natural gradient experiment on solute transport in a sand aquifer. 1. approach and overview of plume movement: *Water Resources Res.*, v. 22, no. 13, p. 2017–2030.
- Mantoglou, A., and Wilson, J. L., 1982. The turning bands methods for simulation of random fields using line generation by a spectral method: *Water Resources Res.*, v. 18, no. 5, p. 1379–1394.
- Masry, E., 1971. Random sampling and reconstruction of spectra: *Information and Control*, v. 19, p. 275–288.
- Matheron, G., 1973. The intrinsic random functions and their applications: *Advan. Appl. Prob.*, v. 5, p. 439–468.
- Matheron, G., and deMarsily, G., 1980. Is transport always diffusive? a counterexample: *Water Resources Res.*, v. 16, no. 5, p. 901–917.
- Neuman, S. P., Winter, C. L., and Newman, C. M., 1987. Stochastic theory of field-scale Fickian Dispersion in anisotropic porous media: *Water Resources Res.*, v. 23, no. 3, p. 453–466.
- Ripley, B. D., 1981. *Spatial statistics*: Wiley-Interscience, New York, 252 p.
- Shinozuka, M., and Jan, C. M., 1972. Digital simulation of random processes and its applications: *Jour. Sound Vib.*, v. 25, no. 5, p. 111–128.
- Smith, L., and Freeze, R. A., 1979. Stochastic analysis of steady state groundwater flow in a bounded domain. two dimensional simulations: *Water Resources Res.*, v. 15, no. 6, p. 1543–1559.