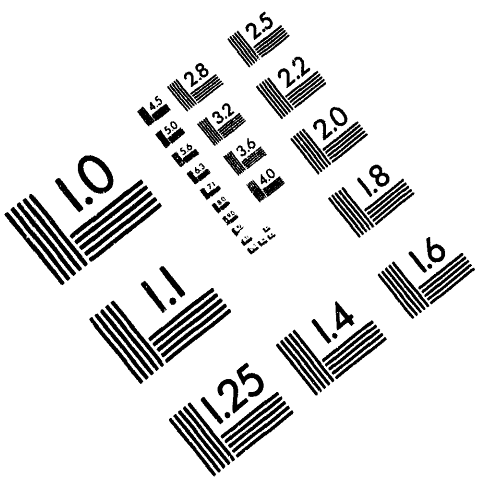
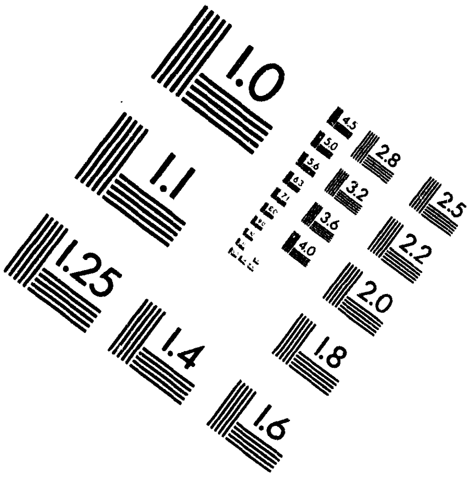


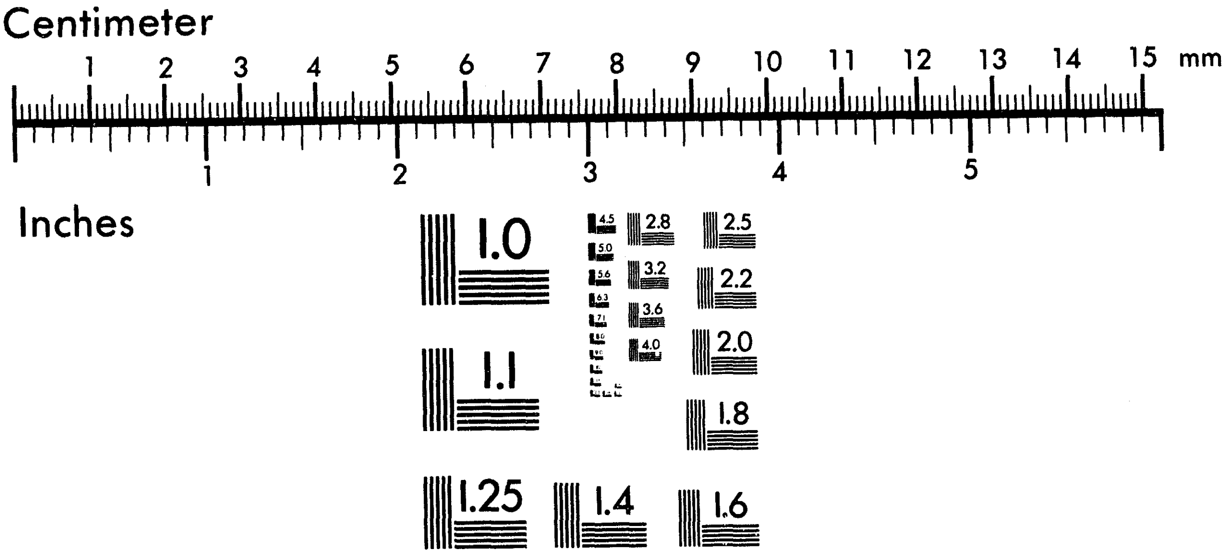
AIIM

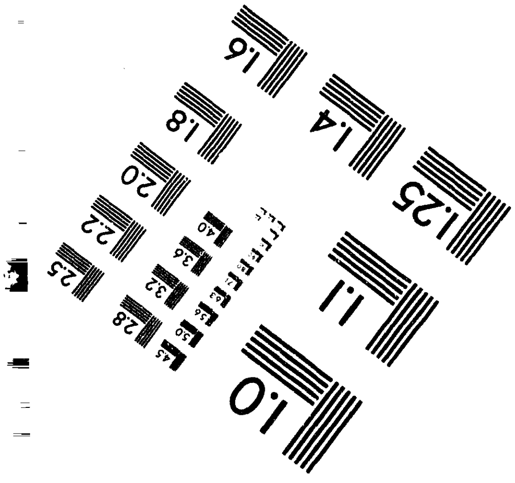
Association for Information and Image Management

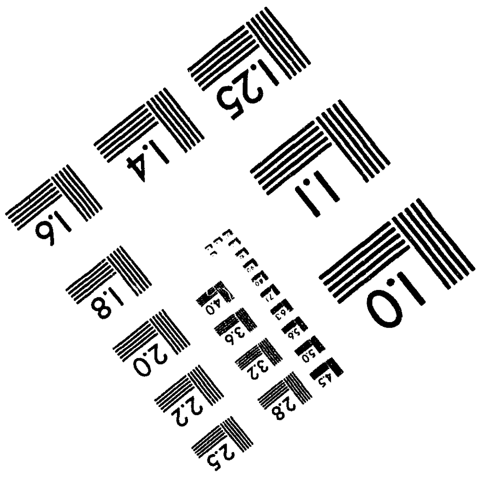
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202









MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

Conf-940748--81

SAND94-2051C

SELECTING FEATURES FROM SPATIAL DATA FOR USE IN STOCHASTIC SIMULATION

Brian M. Rutherford
Statistics and Human Factors Department, Sandia National Laboratories
Albuquerque, NM, 87185-0829

Carol A. Gotway
Department of Biometry, 103 Miller Hall, University of Nebraska
Lincoln, NE, 68583-0712

ABSTRACT

An assessment of the long term containment capabilities of a possible nuclear waste disposal site requires both an understanding of the hydrogeology of the region under consideration and an assessment of the uncertainties associated with this understanding. Stochastic simulation -- the generation of random "realizations" of the regions hydrogeology, consistent with the available information, provides a way to incorporate various types of uncertainty into a prediction of a complex system response such as site containment capability. One statistical problem in stochastic simulation is: What features of the data should be "mimicked" in the realizations? The answer can depend on the application. A discussion is provided of some of the more common data features used in recent applications. These features include spatial covariance functions and measures of the connectivity of extreme values, as examples. Trends and new directions in this area are summarized including a brief description of some statistics (the features) presently in experimental stages.

INTRODUCTION

Site characterization is an important aspect of many problems in waste management and environmental restoration. Often, a groundwater flow model is developed to estimate groundwater travel time to help determine the hazards associated with waste emplacement or a strategy for cleanup at a contaminated site. The flow model is computed by executing software that generally requires a complete description of the hydrogeology in the region -- information that is not available and must be estimated from data usually consisting of bore hole, groundwater and/or soil sample analytic results from within the region or from similar regions. It is this step of going from sample information to a complete specification of the regions hydrogeology that requires geostatistical techniques.

Stochastic simulation is one approach to the inference step mentioned above, however other geostatistical techniques are available. Most alternative approaches rely on providing an estimate (usually a smoothed estimate based on available data) of the hydrogeological parameter values and trying to propagate the related estimation uncertainty through the flow code to obtain an estimate of uncertainty associated with the flow model. Two problems are common when using this approach. First, for complex "transfer functions" like a flow code it is difficult to accurately propagate the uncertainty associated with the estimated surface through the transfer function to obtain an estimate of flow path uncertainty. Second, it is generally accepted that a smoothed surface will not adequately reflect the spatial variability in the hydrogeologic parameters of interest, and hence, this approach may not yield an accurate estimate of the system response.

Stochastic simulation is an alternative approach that avoids these problems. The objective of the simulations is to generate two- or three-dimensional fields that share certain features or properties with the hydrogeologic region of interest. These realizations are then processed through the transfer function to provide an estimate of the system response. By generating a number of realizations and repeating the analysis for each, an uncertainty distribution for the system response can be estimated. For the present example, one would generate maps defining the hydrogeology of the region (the realizations), and use groundwater flow and transport codes (the transfer function) to obtain the related flow model on which the groundwater travel times (the system response) are based. See Figure 1 for a diagram of the general procedure.

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

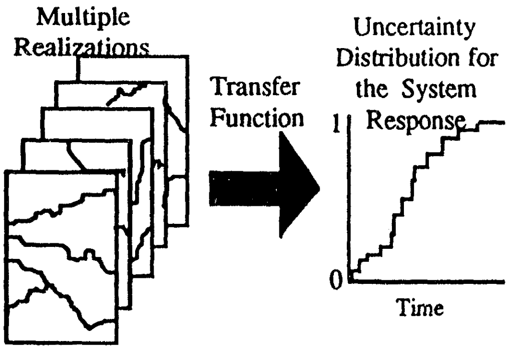


Figure 1. General Procedure for Stochastic Simulation

There are many different simulation approaches that can be used to construct realizations. The approaches differ in a number of ways, including the features of the region that are to be modeled, how closely the generated realizations are constructed to mimic these features and how conditioning information (data, noisy data, or other information based on site knowledge) is utilized. A comparison of several of the most frequently used simulation methods including their description, assumptions and a simulation study comparing their precision and accuracy is given in Gotway and Rutherford (1994). In the present paper, we focus on the first of these differences between approaches -- the features to be modeled. The determination of statistics that should be calculated from sample data, modeled, and input to the simulation algorithm is an important concern in most spatial applications, often complicated by the processing of the realizations through a nonlinear transfer function.

An illustration of differences in the generated realizations that might result from the use of models based on different features of the data is given in Figure 2. Figure 2a shows an image of a trench wall where the rectangular grid points (black or white) indicate the type of geologic material at that location. The trench wall was analyzed as part of the Grater Confinement Disposal (GCD) Project where flow times through the region were dependent on the patterns because of differences in the conductivity of the two types of materials. Figures 2b through 2d show realizations mimicking this trench wall that are generated using different features of the original data and different simulation methods. The differences in features of the realizations may have very different impact on the transfer function when compared to one another. More importantly, any one of the figures may provide a better representation of the properties or features at the site that are important to flow there and hence more accurately predict the containment or environmental restoration capabilities of the site. Different transfer functions will respond differently to alternative sets of spatial features modeled from the data. For any particular application this relationship must be evaluated.

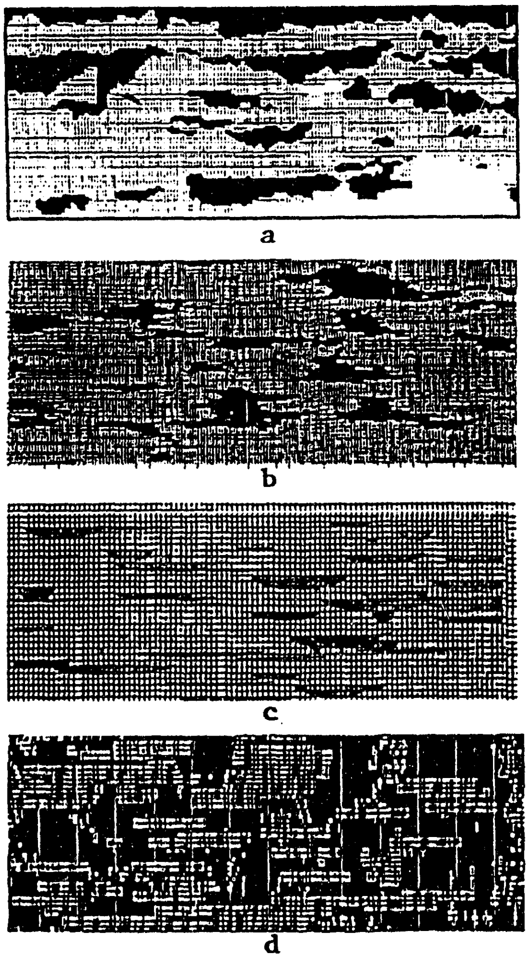


Figure 2. GCD trench wall and realizations based on different features of GCD data.

In the remainder of this paper, we address the questions: 1) What are some of the stochastic models that might be used to represent the hydrogeological region of interest? 2) What features of a region have been modeled in recent applications? and how do they compare? 3) What are some research directions for choosing features for use in stochastic simulation? Answers to these three questions will be discussed for three general characterization problems described in the following section.

STOCHASTIC MCD.3LS

The types of stochastic models that might be useful differ from application to application. The statistical framework associated with most site characterization problems can be put in one of the general categories listed below. These categories are not unique. Many problems (including the GCD example above) can be put loosely into two of the categories. Some of the methods discussed within a category can be, and have been applied in other categories. The categories are established only to try to illustrate basic differences in characterization problems and to simplify the presentation.

A) The region can be represented by a homogeneous random field model. This type of model is likely to apply to the transmissivity field at the Waste Isolation Pilot Project underground site where the entire sampled stratigraphic unit was created through the same depositional and post-depositional processes.

B) The region can be thought of as homogeneous in terms of its large scale geological features, but, smaller features like the orientation and location of fractures or the mixture of different geological materials are the features that are important to flow in the region. These types of stochastic models are likely to be important to flow within a stratigraphic unit at Yucca Mountain that lie below the site of the potential repository for high level nuclear waste or to the GCD Program where the flow model is determined by the relative sizes and locations of two different types of geological materials.

C) The region cannot be thought of as homogeneous in its large scale features. Two different situations can lead to these types of site characterization models. First, distinct regions of different origin may result in completely different hydrogeological characteristics. For situations of this type, the simulation must predict the location size and shape of regions of different geologic materials based on sample information. This is one of the problems likely to be encountered when different stratigraphic units are involved. Another situation that might lead to these types of characterization problems is when portions of the region of interest are much closer to a geological event (the source of volcanic activity, for example), than others. In this case it may be necessary to estimate a trend in the region.

The final three sections of this paper address the questions listed above for these characterization problems. We reference a number of possible approaches but relatively few of these have been applied to problems in stochastic simulation. The applications of and the theory behind the use of random field models for making site specific inferences lies substantially ahead of the use of these same models for stochastic simulation applications. Part of the problem is the lack of a general framework within which to model an arbitrary set of features, understood only through site information, and mimic them in the generated realizations in a way that reflects an appropriate level of understanding. We don't address this problem here -- (differing) opinions on how this process might be accomplished (for the problem of site inference or for simulation) can be found in the literature. Handcock and Stein (1993), the discussion of and rejoinder to Gotway (1994), and Deutsch and Journel (1992b) are recent examples.

HOMOGENEOUS RANDOM FIELD CHARACTERIZATION

Random field models for continuous valued parameters are constructed by taking advantage of spatial relationships among parameter values at neighboring (and relatively close) points. Alternative models include Gaussian and Markov random fields where the latter may assume one of a number of conditional distributions describing the dependence of the value at one point on values at other points within a fixed neighborhood. Gaussian based approaches have dominated the geostatistical literature until just recently. Commonly used methods include LU decomposition, turning bands (Journel 1974), the sequential Gaussian approach (Deutsch and Journel, 1992a) and spectral approaches such as those developed in Mejia and Rodriguez-Iturbe (1974) and Gutjahr (1989). All of these methods use variogram or covariance information modeled from the data as input features but differ in the way they use this information to generate the realizations.

Another approach to simulating continuous random fields is to partition the range of hydrogeological values into discrete categories (possibly representing different spatial relationships) and then generate realizations of discrete-valued fields. The discrete fields are then transformed back to the continuous scale. It was differences in the spatial relationships at different levels of the hydrogeologic parameters that provided one of the primary motivations for partitioning the range of possible parameter values into an ordered set of discrete categories and allowing different (indicator) covariance relationships for each category (Journel and Alabert (1989)). This sequential indicator method uses these indicator variograms to compute probabilities of falling into each category as the algorithm sequentially assigns values to different grid nodes and uses these probabilities to randomly select a category. The categorical assignments are later transformed back to the continuous scale in most applications.

While progress has been made on characterization of Markov random fields (for example Besag, 1974 or Cressie, 1991, Chapters 6 and 7) none of the recent geostatistical simulation applications exploit this statistical framework. Goutsias (1994) describes a method for generating Markov field realizations and investigates several of the theoretical aspects of this approach. The use of methods exploiting other features of the data are being investigated. Experimental work is underway at Sandia National Laboratories with a nonparametric approach where values at different points of the region of interest are simulated sequentially by comparing the pattern observed at the time of simulation to the spatial pattern of the original data. Weights for the relative likelihood of taking on values over any given interval can then be established and the value at that point generated randomly. This nearest neighbor approach requires the definition of metrics that can be used to quantify the closeness of patterns in a reasonable way.

RELATIVELY SMALL SCALE FEATURE CHARACTERIZATION

Regions that are homogenous in terms of large scale features but where flow properties depend on smaller scale features such as a fracture network or the mixture and pattern of different geological materials, for example, require different methods of simulation, in general, than do stationary regions of continuous valued parameters. Very often the data used to characterize such a region is taken from drifts or trenches where a two dimensional subsection of the region can be observed. These subsections are referred to as "training sets" by some authors.

There are several approaches to characterization of these discrete features. One approach is to establish the parameters associated with a point process and to establish distributions for the shape, size and orientation of the geological features of interest. Boolean models and generalizations including marked point processes are examples of this approach used in a number of applications. Another approach is to model the region as a random set characterized by its frequency of intersection with user defined sets or "structuring elements" of various size and shape. A third approach, applicable to the modeling of a fracture network in particular, is to assume a model of random tessellations that adequately matches the training set and retain various features that describe the tessellation model. A detailed treatment of these and other approaches to random set characterization is given in Stoyan et al. (1989).

There have been a number of applications requiring the characterization of small scale features. Modeling fracture networks has been predominate in the geostatistical applications. Chiles (1988) provides an example of two of the approaches described above applied to a granite massif. Corresponding to the Boolean model, Chiles generated a set of "parent" locations using a fixed density Poisson field. He then used a variogram based on fracture occurrence relative to the parent to establish a "parent-daughter" model that captures the clustering of fractures. The point process is "marked" with orientation and trace length values selected at random from their estimated distributions. Chiles also uses a fractal approach to simulate the fracture network where the "local fractal dimension" is calculated using the frequency of intersection of the fracture network in the training set with squares of various size. As such, the analysis is an application of the random set methodology mentioned second above. In a number of applications where the fractal approach has been used, the frequency of intersection approach leads to the definition of a fixed fractal dimension through a certain range of square size. These networks are modeled as self similar fractal networks. Chiles discusses methods for simulation using either set of spatial features.

Tessellation models have been investigated in a number of fracture characterization applications. Gray et al. (1976) relate the fraction of polygons having k vertices or k straight sides and other characteristics of the network to features like percentages of "T", "Y" and "X" joints in the network. Examples in this paper and in Stoyan et al. (1989) Chapter 10, where features such as node intensity and edge midpoint intensity are used, show networks generated by simulating physical processes such as crack growth or crystal growth from an initial point process or sequentially generated point process. Simulations conducted by generating random hydrogeological processes in a region are available, but not common in geostatistical applications.

The use of methods exploiting other features of the data are under investigation. Guardiano and Srivastava (1992) use multi-point features of the data to model a number of complex geometries using a method they call extended normal equations algorithm. This algorithm is like a discrete version of the nearest neighbor approach where the metric defining closeness of patterns is provided by Bayes Theorem calculated using the frequency of occurrences of the patterns in the training data set. Stoyan et al. (1989) describes another approach to modeling discrete features that might prove useful particularly when the shape of the discrete features is of importance. This method requires that the discrete features be mapped to an appropriate "representation space" where they can be modeled as realizations of a point process. A directed line process in two dimensions, for example, can be modeled as a point process on a cylinder in three-dimensional space.

LARGE SCALE FEATURE CHARACTERIZATION

One problem with using a single set of features to specify spatial relationships for an entire region is that this set of features is then applicable to all values over the range of the hydrogeologic parameter to be simulated. This may prove inappropriate if: (1) there are distinct subsets of the region where the general pattern of the spatial relationship does not apply or (2) there are significant trends in the parameter values throughout the region. For applications where (1) applies, an appropriate approach requires, first, trying to establish the thickness or shape of specific subsets of the region (different stratigraphic units for example), and later, simulating values within these sub-regions. The former step leads to categorical or discrete variable simulations which, in general, involve a different set of assumptions and different stochastic simulation methods then continuous parameter simulations.

The two techniques encountered most in geostatistical applications are the truncated Gaussian approach (Galli et al. 1990 and Dowd, 1992) and a discrete version of the sequential indicator approach (Deutsch and Journel 1992a). The former method is restricted to a single covariance function for expressing the continuity within sub-regions and the spatial relationships between sub-regions. The latter method permits an indicator variogram to be specified for each sub-region but provides no guidance for between sub-region relationships beyond that given in conditioning data (conditioning sample data will help order the sub-regions in a meaningful way that reflects site information). A comparison of these two approaches is included in Gotway and Rutherford (1994).

Some of the approaches described in previous subsections are applicable to large scale feature characterization. Nearest neighbor and multipoint covariance function approaches may prove useful in this type of application. For cases where the specific shape of the sub-regions is important, methods used for the characterization of random sets may be more appropriate.

For applications where (2) applies, an unanswered question in the geostatistics literature is whether trends in the hydrogeological parameters of interest should be modeled explicitly first, and features of the residuals should be used in the analysis, or whether the trends can be adequately captured through using features of the original data. A third alternative is to model a region with trend as an intrinsic random function with stationary increments. The geostatistical literature is rich in alternative methods for modeling, particularly when Gaussian random fields with a trend are involved. Methods and applications of stochastic simulation by comparison are very sparse for this situation. In many cases the trend is ignored. Methods are applied even though some of the assumptions may not be satisfied.

The use of methods exploiting models with trend are being investigated. Armstrong (1991) describes the progress in simulating random fields with stationary increments using generalized covariance models. Universal kriging or alternative methods that separate large and small scale variability such as median polish kriging (Cressie (1986)) can be used as long as estimates of the uncertainty in the trend parameters are available along with estimates of the covariability of the de-trended random field. The simulation would then proceed -- generate a trend at random, then add to this a realization of the de-trended field representing small scale variability. Features of the de-trended field may or may not have to be recomputed after the trend has been generated. A comparison of the utility of these approaches in stochastic simulation would be useful.

REFERENCES

Armstrong, M. (1991). Personal correspondence to Carol Gotway dated 5/15/91.

Besag, J. E. (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society B*, 36, 192-225.

Cressie, N. (1991). *Statistics for Spatial Data*, John Wiley and Sons, Inc., New York.

Cressie, N. (1986). "Kriging Nonstationary Data" *Journal of the American Statistical Association*, 81, 625-634.

Chiles, J. P. (1988). "Fractal and Geostatistical Methods for Modeling of a Fracture Network," *Mathematical Geology*, 20, 631-654.

Deutsch, C. V. and Journel, A. G. (1992a). *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York.

Deutsch, C. V. and Journel, A. G. (1992b). "Annealing Techniques Applied to the Integration of Geological and Engineering Data," *Stanford Center for Reservoir Forecasting Annual report 5*, Stanford University, Stanford Ca.

Dowd, P. A. (1992). "A Review of Recent Developments in Geostatistics," *Computers and Geosciences*, 17, 1481-1500.

Galli, A., Guerillot, D., Ravenne, C. and HERESIM Group (1990). "Combining Geology, Geostatistics, and Multiphase Fluid Flow for 3D Reservoir Studies," In *Proceedings of the 2nd European Conference on the Mathematics of Oil Recovery*, Guerillot, D., and Guillon, O. eds. Editions Technip, Paris, France, 11-19.

Gotway, C. A. and Rutherford, B. M. (1993). "Stochastic Simulation for Imaging Spatial Uncertainty: Comparison and Evaluation of Available Algorithms," *Geostatistical Simulations*, Armstrong M. and Dowd P. eds. Kluwer Academic Publishers, The Netherlands..

Gotway, C. A. (1994). "The Use of Conditional Simulation in Nuclear-Waste-Site Performance Assessment," *Technometrics*, 36, 129-140.

Goutsias, J. (1994). "A Theoretical Analysis of Monte Carlo Algorithms for the Simulation of Gibbs Random Field Images," IEEE Transactions on Information Theory, 37, 1618-1628.

Gray, N. H., Anderson, J. B., Devine, J. D. and Kwasnik, J. M. (1976). "Topological Properties of Random Crack Networks," *Mathematical Geology*, 8, 617-626.

Guardiano, F. B. and Srivastava, M. R. (1992). "Borrowing Complex Geometries From Training Images: The extended Normal Equations Algorithm," *Stanford Center for Reservoir Forecasting Annual report 5*, Stanford University, Stanford Ca.

Gutjahr, A. L. (1989). "Fast Fourier Transforms for Random Field Generation," Project Report for Los Alamos Grant to New Mexico Tech. Contract Number 4-R58-2690R.

Handcock M. and Stein M. (1994). "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403-410.

Journel, A. G. (1974). "Geostatistics for Conditional Simulation of Ore Bodies," *Economic Geology*, 69, 673-687.

Journel, A. G. and Alabert, F. (1989). "Non-Gaussian Data Expansion in the Earth Sciences," *Terra Nova*, 1, 123-134.

Mejia, J. and Rodriguez-Iturbe, I. (1974). "On the Synthesis of Random Fields from the Spectrum: An Application to the Generation of Hydrologic Spatial Processes," *Water Resources Research*, 10, 705-711.

Stoyan, D., Kendall, W. S. and Mecke J. (1987). *Stochastic Geometry and Its Applications*, Wiley, New York and Akademie-Verlag, Berlin.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DATE
FILMED
9/19/94

END