

Conditions For Frequency Domain Multi-Gaussian Simulation
With A Review of the Theory of Frequency Domain Conditional Simulation

John W. Kern¹

Leon E. Borgman²

Daniel Goodman³

¹ Kern Statistical Services, Inc, 415 NW Robert, Pullman WA 99163.

² Departments of Statistics, and Geology and Geophysics, University of Wyoming, Laramie WY.
82071

³ Montana State University, Department of Biology, Bozeman MT.

RH: Consistency Conditions for Frequency Domain Simulation

Correspondence to:

John W. Kern

415 NW Robert

Pullman WA 99163

Tel: 509-339-2489

Fax: 509-339-2490

e-mail: johnkern@west-inc.com

ABSTRACT

Inconsistencies in the geostatistical literature regarding frequency domain, or spectral simulation methods are identified. We point out that the spectral decomposition methods have not been widely applied due to some misconceptions in the geostatistical literature and due to computational difficulties which can accompany naive application of the spectral methods. In hopes of broadening awareness of the geostatistical community regarding the computationally efficient FFT methods, we review existing spectral methods for conditional simulation.

Additionally, we derive 2 new consistency conditions which are required to control the stability of the FFT methods. Enforcing the new consistency conditions into computer algorithms will allow practitioners to apply the FFT methods with confidence that simulated realizations do in fact possess the proper second order statistics and that realizations interpolate observed data as required.

Key Words: Conditional simulation geostatistics ocean waves spectral Fourier transform

INTRODUCTION

Simulation and conditional simulation of spatial data has become widely known in recent years as an alternative to kriging and other interpolation methods when local extreme values are of interest. Much of the literature on conditional simulation has focused on sequential simulation algorithms (Deutsch and Journel 1992) which are flexible and less computationally intensive for large simulated fields than matrix inversion methods such as LU decomposition (Davis 1987). Frequency domain simulation methods have also been proposed for simulation of random fields as early as 1982, (Borgman 1982). Despite the computational efficiency of these algorithms particularly for simulation of random fields over large regular grids, these methods have been much less widely applied or discussed in the geostatistical literature. We believe that there are 2 primary reasons for the lack of acceptance of these algorithms for geological and environmental problems: 1) the algorithms may be sensitive to certain consistency conditions between the data, the data to data covariance matrix, and the data to simulation grid covariance matrices, and 2) a common misconception in the literature has been that these methods cannot be conditioned to observed data.

Algorithms for conditional simulation using the fast Fourier Transform (FFT) were proposed for large 2 space dimension problems by Borgman et al (1984) and later for arbitrary spatial dimension and multiple dimensional observation vectors by Borgman and Faucette (1993a and 1993b). Applications have been published in the geological literature (Easley, et al 1991) and in oceanography (Borgman et al 1994). In spite of these and other theoretical and applied contributions to the literature, recent texts and publications continue to refer to frequency domain methods as primarily applicable to unconditional simulation problems. Cressie (1991) and

Deutch Journal (1992) simply remain silent on the application of multidimensional FFT methods, although Cressie does describe the unconditional algorithm and discusses speed and efficiency issues. More notably Goovaerts (1997) states in a discussion of indicator simulation, '*Much faster simulation algorithms exist for reproduction of any covariance model as long as there is no data-conditioning, for example, simulation using spectral decomposition...*' and Yao (1998) states that '*spectral simulation only generates unconditional realizations*'. To correct this perceived deficiency in frequency domain methods, Yao proposes an iterative technique for '*phase identification*', a term coined for conditioning the simulation to observed data. We argue that introduction of this new term to the literature is unnecessary and is likely to confuse future attempts to follow an already jargon rich literature. We speculate that the iterative steps required in this technique will likely nullify any gains in efficiency which may be presented by the FFT algorithm.

In this paper we present a non-iterative solution to the conditional simulation problem for the spectral decomposition algorithm. Although the algorithm has been published previously, Borgman (1982), and most notably in the geostatistics literature by Borgman et al (1984), Easley et al (1994), we review of the method to provide a basis for the two new results which are presented.

In addition to a review of the frequency domain method, we identify two consistency conditions which are required for a multi-gaussian conditional simulation to interpolate the observed data and for the resulting simulation to be consistent with the theoretical covariance function which is to be reproduced. This result applies to frequency domain methods, sequential Gaussian

algorithms as well as matrix decomposition methods and algorithms which utilize these methods on normal scores transforms of the data. It is shown that not all spatial data configurations are consistent with the theoretical properties of the random field to be simulated. In particular, when these consistency conditions are violated, the iterative scheme proposed by Yao (1998) is not expected to converge.

We present a method to handle this situation in using orthogonal projection of the data vector onto the column space of the data covariance matrix. In particular, this projection scheme is easily implemented with the frequency domain algorithm. In our experience when these consistency conditions are not enforced, conditional simulations may fail to interpolate known data and in cases of severe inconsistency, simulations may be nearly independent of the observed data. In these situations, the simulations may appear like the result of fitting a very stiff spline function to the data. We suspect that the uninitiated practitioner may run into these cases and give up on the FFT methods in favor of the less efficient but more user friendly sequential methods. The consistency conditions which we have proven here allow the user to diagnose and work around these situations. Regardless of the simulation scheme, these diagnostic measures also allow improved assessment of the relationship between the observed data and the theoretical model from which realizations are to be drawn. It is not clear that valid statistical inferences can be made when data are not consistent with the theoretical models upon which inferences are to be based.

Finally, in this paper we present the results of empirical comparisons of the speed and accuracy of the proposed FFT algorithm to the sequential Gaussian algorithm SGSIM, (Deutsch and

Journal, 1992). We do not propose that the frequency domain method is a replacement for other algorithms. Although we do propose that the FFT algorithm may have certain advantages over the sequential algorithms particularly in an environmental regulatory setting where inferences cannot be based on a small number of realizations, and where a complete understanding of the relationship between the theoretical model and the sample data is critical. An applied case study of the application of frequency domain conditional simulation to heavy metals contamination is provided in a subsequent paper.

REVIEW OF FREQUENCY DOMAIN CONDITIONAL SIMULATION

We consulted three recent texts (Deutsch and Journel 1992, Cressie 1991, and Goovaerts 1997) for a definition of conditional simulation and found that none gave a precise definition although all seemed in general agreement that that a conditional simulation should be a set of modeled data which agree with certain aspects of the statistical distribution of observed sample data, and which interpolate (agree with) the observed sample data at the sampling locations. For the multivariate normal case this can be made precise.

Definition: Conditional Simulation

Given a set of sample data collected at known locations, a multivariate normal conditional simulation is any set of numbers which interpolates the known data at a set of sampled locations and has the same second order statistics, mean and covariance as the observed data. In the following development we discuss conditional simulation of multivariate normal random fields. For a large class of non-Gaussian random fields, these models can be applied after a normal scores, indicator or other appropriate transformation.

Conditional Simulation of Multivariate Normal Data

Let

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} \quad \text{with} \quad E(\mathbf{W}) = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad (1)$$

be a multivariate normally distributed random vector, where \mathbf{W} is understood to be indexed to a collection of locations in space, and \mathbf{W}_1 represents the random variable at a set of known sampling locations and \mathbf{W}_2 represents a set of locations at which simulated values are to be produced (e.g. these are usually the nodes of a regular grid). A conditional simulation of \mathbf{W}_2 given $\mathbf{W}_1 = \mathbf{w}_1$ is given by

$$\mathbf{W}_{cs} = C_{12}^T C_{11}^{-1} (\mathbf{w}_1 - \mathbf{W}_{1us}) + \mathbf{W}_{us}, \quad (2)$$

where \mathbf{W}_{us} is an unconditional simulation at the desired spatial locations. and \mathbf{W}_1 is the unconditional simulation at the known points.

If \mathbf{W}_1 and \mathbf{W}_{us} are multivariate normally distributed, then \mathbf{W}_{cs} is multivariate normally distributed with conditional mean and covariance (Borgman, 1982),

$$E(\mathbf{W}_{cs}) = C_{12}^T C_{11}^{-1} (\mathbf{w}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_{us}, \quad (3)$$

and

$$\text{cov}(\mathbf{W}_{cs}) = C_{22} - C_{12}^T C_{11}^{-1} C_{12}. \quad (4)$$

Provided that all of the matrices are of reasonable size, this matrix simulation method is efficient and straight forward. However, when large spatial grids are to be simulated, these methods

become computationally prohibitive, even when sparse matrix methods are used. For example, when the spatial grid is 1000 by 1000, the matrix inversion required to produce W_{us} is 1,000,000 by 1,000,000. This computational burden motivated the development of both the sequential simulation and the FFT methods.

Frequency Domain Simulation

The utility of the frequency domain methods is that the FFT can be used to transform the space domain random field with long range spatial correlations and constant variance into a frequency domain function with components that are statistically independent and which have frequency dependent non-constant variance. The following development of the frequency domain methods is a synthesis of papers by Borgman (1982), Borgman, et al(1984), and Borgman and Faucette, (1993a and 1993b).

If $w(\mathbf{x})$ is a square integrable function

$$\int_{-\infty}^{\infty} |w(\mathbf{x})|^2 d\mathbf{x} < \infty, \quad (5)$$

there exists a function $A(\mathbf{f})$ such that

$$A(\mathbf{f}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w(\mathbf{x}) e^{-i2\pi\mathbf{f}^T\mathbf{x}} d\mathbf{x}, \quad (6)$$

and

$$w(\mathbf{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(\mathbf{f}) e^{i2\pi\mathbf{f}^T\mathbf{x}} d\mathbf{f} \quad (7)$$

where $i = \sqrt{-1}$. We say $w(\mathbf{x})$ and $A(\mathbf{f})$ are Fourier transform pairs.

The covariance function of a stationary random function also has a frequency domain Fourier pair called the spectral density function or spectrum

$$S(\mathbf{f}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(\mathbf{h}) e^{-i2\pi\mathbf{f}^T\mathbf{h}} d\mathbf{h}. \quad (8)$$

The spectrum can be interpreted as the variance associated with oscillations at a particular frequency in the space domain.

Define $w(n_1, n_2)$ to be a discrete 2 dimensional field such that $w(x_1, x_2) = w(n_1\Delta x_1, n_2\Delta x_2)$ for $n_1 = -N_1/2, -N_1/2+1, \dots, N_1/2$, and $n_2 = -N_2/2, -N_2/2+1, \dots, N_2/2$. Let $T_1 = N_1\Delta x_1$, $T_2 = N_2\Delta x_2$, and let $f_1 = m_1\Delta f_1$ and $f_2 = m_2\Delta f_2$ such that

$$f_1 x_1 = \frac{m_1 n_1}{(\Delta f_1 \Delta x_1)^{-1}}, \quad \text{and} \quad f_2 x_2 = \frac{m_2 n_2}{(\Delta f_2 \Delta x_2)^{-1}}. \quad (9)$$

Now if Δf_1 and Δf_2 are selected so that $(\Delta f_i \Delta x_i)^{-1} = N_i$, then $f_i x_i = m_i n_i / N_i$ and $\Delta f_i = 1/T_i$.

With these conventions, the discrete analogues to equations 6 and 7 can be written as

$$A(m_1, m_2) = \Delta x_1 \Delta x_2 \sum_{n_1 = -N_1/2}^{N_1/2} \sum_{n_2 = -N_2/2}^{N_2/2} W(n_1, n_2) \exp \left[-i2\pi \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_2}{N_2} \right) \right] \quad (10)$$

and

$$W(n_1, n_2) = \Delta f_1 \Delta f_2 \sum_{m_1=-N_1/2}^{N_1/2} \sum_{m_2=-N_2/2}^{N_2/2} A(m_1, m_2) \exp \left[+i2\pi \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_2}{N_2} \right) \right]. \quad (11)$$

To take advantage of the computational efficiencies of the FFT, the sequence $W(n_1, n_2)$ is assumed to be periodic with period T_1 and T_2 in each direction. This is equivalent to treating the simulation grid as a window on one cycle of a periodic sequence. Periodicity within the study area is not needed which is a common misconception regarding frequency domain methods. This assumed periodicity also has the effect of allowing one to shift the indices to the set $n_1 = 0, 1, \dots, N_1-1$ and $n_2 = 0, 1, \dots, N_2-1$. This is the convention used by common FFT algorithms and will be used here unless otherwise noted. In this notation, equation 10 can be restated as

$$A(m_1, m_2) = \Delta x_1 \Delta x_2 \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} W(n_1, n_2) \exp \left[-i2\pi \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_2}{N_2} \right) \right] \quad (12)$$

and the other formulas are similarly modified. In the discrete coordinates the spectrum and covariance functions become

$$S(m_1, m_2) = \Delta x_1 \Delta x_2 \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} C(n_1, n_2) \exp \left[-i2\pi \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_2}{N_2} \right) \right] \quad (13)$$

and

$$C(n_1, n_2) = \Delta f_1 \Delta f_2 \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} S(m_1, m_2) \exp \left[+i2\pi \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_2}{N_2} \right) \right]. \quad (14)$$

In general, if W is real, A will be complex valued, and can be written in the form

$$A_{m_1 m_2} = U_{m_1 m_2} + iV_{m_1 m_2}, \text{ where } U \text{ and } V \text{ are the real and complex parts of } A \text{ respectively.}$$

The utility of the frequency domain methods hinges on the fact that if W is multivariate normal with mean zero, the real and complex parts of $A(m_1, m_2)$ are normally distributed with mean zero and are uncorrelated within certain zones of the frequency domain and related by complex conjugation in other zones (Taheri, 1980). The variances of the real and complex parts of A are given in Table 1 for the zones (A, B, C₁ and C₂) defined in Figure 1. These relations allow simulation of mean zero independent random variates in each zone with the appropriate frequency dependent variance. The remaining array of Fourier coefficients is filled in through the symmetries

$$A_{N_1-m_1, N_2-m_2} = \bar{A}_{m_1, m_2}, \quad A_{N_1-m_1, m_2} = \bar{A}_{m_1, N_2-m_2}. \quad (15)$$

for $0 < m_1 < N_1/2$ and $0 < m_2 < N_2/2$.

Table 1. Variance covariance relations for real and complex parts of the Fourier coefficients, $A(m_1, m_2)$.

$var(U_{m_1 m_2})$	$var(V_{m_1 m_2})$	$cov(U_{m_1 m_2}, V_{m_1 m_2})$	m_1	m_2	Zone
$T_1 T_2 S_{m_1 m_2}$	0	0	0	0	A
$T_1 T_2 S_{m_1 m_2}$	0	0	0	$N_2/2$	A
$T_1 T_2 S_{m_1 m_2}$	0	0	$N_1/2$	0	A
$T_1 T_2 S_{m_1 m_2}$	0	0	$N_1/2$	$N_2/2$	A
$0.5 T_1 T_2 S_{m_1 m_2}$	$0.5 T_1 T_2 S_{m_1 m_2}$	0	0	$1 : N_2/2$	B
$0.5 T_1 T_2 S_{m_1 m_2}$	$0.5 T_1 T_2 S_{m_1 m_2}$	0	$N_1/2$	$1 : N_2/2$	B
$0.5 T_1 T_2 S_{m_1 m_2}$	$0.5 T_1 T_2 S_{m_1 m_2}$	0	$1 : N_1/2$	0	B
$0.5 T_1 T_2 S_{m_1 m_2}$	$0.5 T_1 T_2 S_{m_1 m_2}$	0	$1 : N_1/2$	$N_2/2$	B
$0.5 T_1 T_2 S_{m_1 m_2}$	$0.5 T_1 T_2 S_{m_1 m_2}$	0	$1 : N_1/2-1$	$1 : N_2/2-1$	C_1
$0.5 T_1 T_2 S_{m_1 N_2 - m_2}$	$0.5 T_1 T_2 S_{m_1 N_2 - m_2}$	0	$1 : N_1/2-1$	$N_2/2+1:N_2-1$	C_2

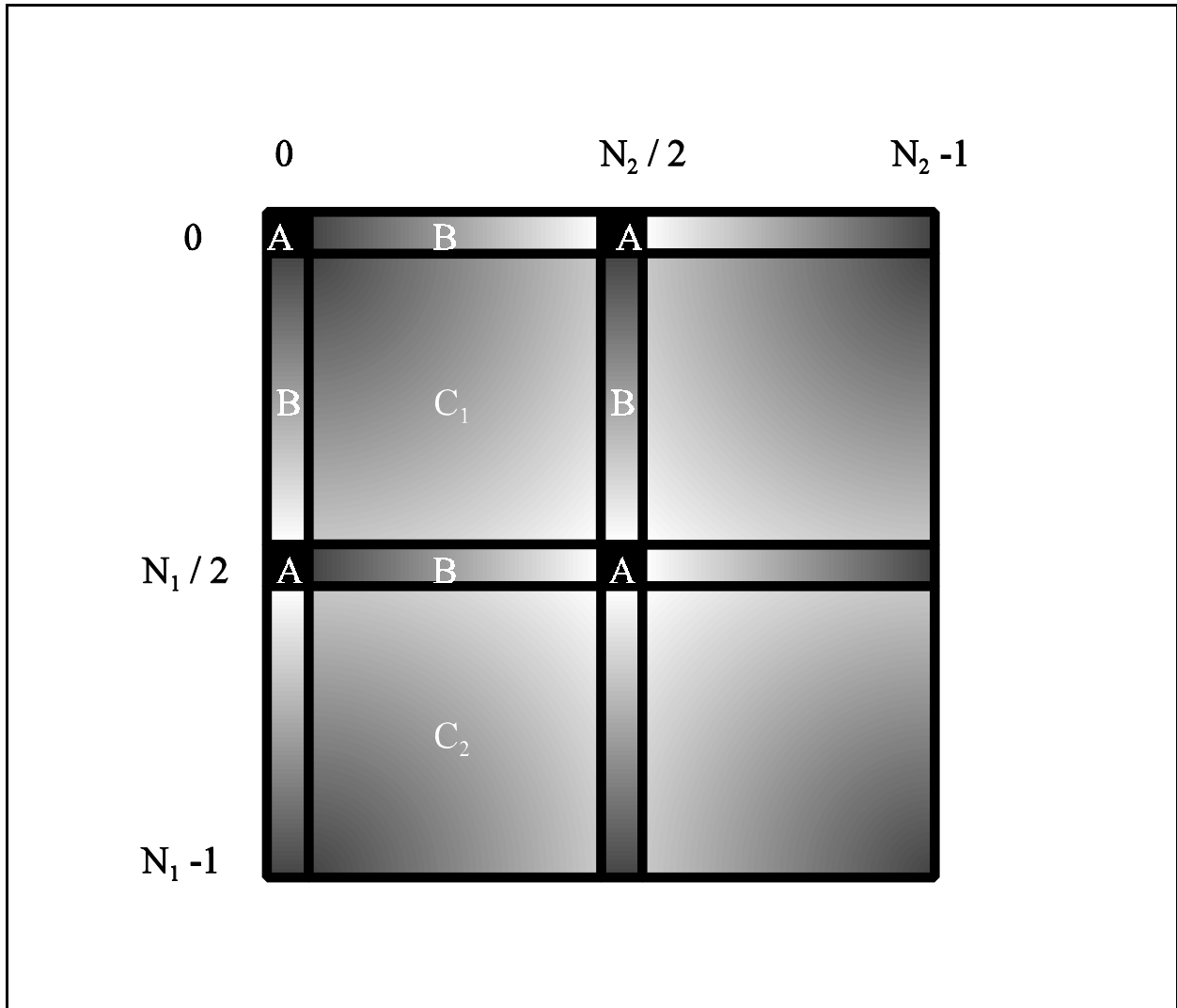


Figure 1. Organization of real and imaginary parts of the Fourier coefficients. Zones referred to in Table 1 are labeled and required symmetries are indicated by the gradient in fill color. This bookkeeping system is referred to as the windowed form of the Fourier coefficients. These symmetries are required to take advantage of the fast Fourier transform algorithm FT235 (Borgman and Yfantis 1981) .

Based on these relationships, an unconditional simulation can be obtained on an N_1 by N_2 array of locations by simulating N_1N_2 independent standard Normal random variables in frequency domain with the correct variance, followed by inverse Fourier transforming the simulated frequency domain variates.

Unconditional Simulation Algorithm

- 1) Estimate the mean and covariance function from data, or simply choose a desired covariance model.
- 2) Obtain corresponding spectrum $S(m_1, m_2)$ either analytically or by FFT of the 2-D covariance model.
- 3) Simulate N_1N_2 independent normally distributed variates with variance as a function of $S(m_1, m_2)$ within zones A, B, C_1 , and C_2 .
- 4) Obtain remaining Fourier coefficients by symmetries and complex conjugation.
- 5) Inverse Fourier transform the simulated coefficients using 2-D FFT.

Conditioning to Sample Data

The simulation can be conditioned (e.g. made to interpolate) the known data using

$$W_{cs} = C_{I2}^T C_{II}^{-1} (w_1 - W_1) + W_{us}. \quad (16)$$

This is the same technique used in the matrix decomposition method with the exception that the unconditional simulation comes from the frequency domain approach. It should be noted that this method requires one matrix inversion and one matrix multiplication to produce the first realization, and that successive realizations are produced with a single FFT and a single

matrix multiplication. At this step Yao (1998) applies an iterative scheme to approximate W_{cs} . When multiple realizations are desired, and the number of conditioning points is moderate (e.g. 10 to a few hundred) equation 16 will be more efficient than Yao's method or sequential methods (Journel and Huijbregts 1978). The primary limitation on the technique is that the number of data observations must be reasonable relative to the available computing resources.

CONSISTENCY CONDITIONS

In practice there are cases in which a conditional simulation using the frequency domain method does not precisely interpolate the sample data. It will be shown in this section that this can result from inconsistencies between the observed data and the covariance matrix C_{11} , or by inconsistencies between C_{11} and C_{12} . We believe that this may have been a deterrent to the widespread adoption of frequency domain methods as unpredictable results can occur if the appropriate consistency conditions are not enforced. It should be noted that these consistency requirements hold equally for space domain algorithms. However, the sequential Gaussian methods produce results even when there are inconsistencies between data and theoretical models. In the following sections, we present two new results which allow practitioners to apply the frequency domain methods with confidence that the resulting simulations accurately reproduce second order properties of the data and do in fact interpolate the data as desired. Additionally, these results provide diagnostics to investigate the consistency between data and the statistical model being applied.

In order to investigate the needed consistency conditions, the simulation algorithm is re-defined entirely in frequency domain and it is shown that constraints can be placed on C_{12} and C_{11} to guarantee correct interpolation of the observed data. Without the consistency constraints, the

simulation may simply result in a least squares approximation to the conditional simulation and may not fit the data well.

Define A_{us} to be a complex vector of length N_1N_2 of stacked real and complex Fourier coefficients

$$A_{us} = \text{stack}(A) = \begin{bmatrix} U(:,1) - iV(:,1) \\ U(:,2) - iV(:,2) \\ \vdots \\ U(:,N_2) - iV(:,N_2) \end{bmatrix}, \quad (17)$$

where $U(:,j) = [U(1,j), U(2,j), \dots, U(N_1,j)]^T$. Further define the N_1N_2 by N_1N_2 matrix F^* such that

$$F^* A_{us} = \text{stack} \left(\Delta f_1 \Delta f_2 \sum_{m_1=-N_1/2}^{N_1/2} \sum_{m_2=-N_2/2}^{N_2/2} A(m_1, m_2) \exp \left[+i2\pi \left(\frac{m_1 n_1}{N_1} + \frac{m_2 n_2}{N_2} \right) \right] \right), \quad (18)$$

for $-N_1/2 \leq m_1 \leq N_1/2$, and $-N_2/2 \leq m_2 \leq N_2/2$. Defined in this way, the unconditional simulation can be expressed as the matrix product $W_{us} = F^* A_{us}$. It can be shown that (Borgman 1984)

$$A_{cs} = C_{21} C_{11}^{-1} (w_{observed} - W_{us\ observed}) + A_{us} \quad (19)$$

is a conditional simulation in frequency domain if the matrix C_{21} is re-interpreted to represent

$$C_{21} = \text{cov}(A_{us}, W) = E(A_{us} W^H), \quad (20)$$

where $B^H \equiv \bar{B}^T$, the conjugate transpose of B and $W_{us\ observed}$ is a space domain unconditional

simulation at the locations associated with $\mathbf{w}_{\text{observed}}$. The space domain conditional simulation is obtained from the inverse Fourier transform $F^* \mathbf{A}_{cs}$ of the frequency domain conditional simulation.

Theorem I. Consistency Between C_{11} and C_{21}

Define F to be the n by $N_1 N_2$ matrix containing the inverse Fourier transform terms associated with each of the n data locations so that $\mathbf{W}_{us \text{ observed}} = F \mathbf{A}_{us}$, is an unconditional simulation at the data locations with measurements $\mathbf{w}_{\text{observed}}$. If C_{11} is invertible and $C_{11} = F C_{21}$, then $\mathbf{W}_{cs \text{ observed}}$ interpolates the observed data exactly (e.g. the data to data, and the data to simulation covariance matrices are consistent).

Proof:

Express the conditional simulation at the observed data locations as

$$\begin{aligned} \mathbf{W}_{cs \text{ observed}} &= F \mathbf{A}_{cs} \\ &= F C_{21} C_{11}^{-1} (\mathbf{w}_{\text{observed}} - \mathbf{W}_{us \text{ observed}}) + F \mathbf{A}_{us} \\ &= F C_{21} C_{11}^{-1} (\mathbf{w}_{\text{observed}} - \mathbf{W}_{us \text{ observed}}) + \mathbf{W}_{us \text{ observed}} \end{aligned} \quad (21)$$

To guarantee that the conditional simulation interpolates the data, it is sufficient that

$$F C_{21} C_{11}^{-1} = I, \text{ or equivalently } C_{11} = F C_{21}.$$

One interpretation of this relationship is that the data covariance matrix must be the Fourier transform of the covariance between the data and the Fourier coefficients of the unconditional simulation. It should be noted that although this result was derived based on the frequency domain expansion of the simulation algorithm, this result holds for any gaussian

simulation algorithm. Simulated realizations are not a true samples from the multi Gaussian model if these consistencies do not hold.

For the frequency domain approach, this consistency can be built into the computational algorithm if C_{2l} is calculated first from the specified spectral density model, and then C_{1l} is calculated from C_{2l} . The difference between C_{1l} calculated from C_{2l} and the alternative method based directly on the covariance model can be used to diagnose the severity of inconsistency between data and simulation grid covariances. In practice, it is inconsistency between data and the covariance model which tend to have the most severe effects on the simulated realization.

If C_{1l} is full rank, consistency between C_{2l} and C_{1l} is sufficient to guarantee exact interpolation for any observed data vector \mathbf{w} . However, in practice, the modeled covariance matrix C_{1l} may be singular to machine precision (nearly singular). In this situation the required inverse does not exist, and **Theorem 1** does not guarantee that the simulation will interpolate the observed data.

Theorem II. Consistency Between Observed Data and C_{1l}

If the data vector and the unconditional simulation at the data locations are both in the space spanned by the eigenvectors of C_{1l} and C_{1l} is consistent with C_{2l} as in **Theorem 1**, then the conditional simulation exactly interpolates the observed data.

Proof:

The Moore Penrose generalized inverse of C_{1l} is

$$C_{1l}^+ = U_1 L_{1l}^{-1} U_1^T, \quad (22)$$

where U_I is a matrix whose columns are eigenvectors corresponding to the nonzero eigenvalues of C , and L_{II} is a diagonal matrix of non-zero eigenvalues of C_{II} . Let W_{observed} and w_{observed} be in the space spanned by the columns of U_I (e.g. $w_{\text{observed}} = U_I \mathbf{b}$, and $W_{\text{us observed}} = U_I \mathbf{B}$ for some pair of vectors \mathbf{b} and \mathbf{B} respectively). Replacing C_{II}^{-1} with C_{II}^+ in equation (21) and using the orthogonality of the eigenvectors (e.g. $U_I^T U_I = I$) gives

$$\begin{aligned}
W_{\text{cs observed}} &= FC_{2I} C_{II}^+ (U_I \mathbf{b} - U_I \mathbf{B}) + U_I \mathbf{B} \\
&= C_{II} C_{II}^+ (U_I \mathbf{b} - U_I \mathbf{B}) + U_I \mathbf{B} \\
&= U_I L_{II} U_I^T U_I L_{II}^{-1} U_I^T U_I (\mathbf{b} - \mathbf{B}) + U_I \mathbf{B} \\
&= U_I \mathbf{b} = w_{\text{observed}} ,
\end{aligned} \tag{23}$$

the desired result.

In cases where $w \neq U_I \mathbf{b}$, a remedial measure is needed to provide an approximate technique which will guarantee $W_{\text{cs observed}} \cong w_{\text{observed}}$.

One alternative is to condition the simulation to a projection of w_{observed} onto the subspace spanned by the columns of U_I . This amounts to replacing the data vector with it's projection

$$w_p = \text{proj}(w_{\text{observed}}) = U_I U_I^T w_{\text{observed}} \tag{24}$$

The quality of this approximation can be evaluated through an analysis of the residuals

$$\mathbf{r} = (U_I U_I^T - I) w_{\text{observed}} \tag{25}$$

If the maximum absolute error or root mean squared error is large relative to the accuracy needed

for a particular application, then an alternative approach may be explored or the residuals may be used to identify spatial outliers. In our experience, areas of inconsistency are typically noted by observations which differ substantially in value but which are in close spatial proximity.

Another remedial alternative is to perturb the covariance matrix C_{11} slightly to force $\mathbf{w}_{\text{observed}}$ to be in the subspace spanned by the perturbed matrix C_{11}^* . One perturbation could be

achieved through the spectral decomposition $C_{11} \cong \sum_{i=1}^{\nu} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, where ν is the number of

nonzero eigenvalues of C_{11} . When the vector $\mathbf{w}_{\text{observed}}$ is not in the span of C_{11} , C_{11}^* can be

obtained by adding a small multiple of the vector $\mathbf{w}_{\text{observed}}$

$$C_{11}^* = \sum_{i=1}^{\nu} \lambda_i \mathbf{u}_i \mathbf{u}_i^T + \alpha \frac{\mathbf{w} \mathbf{w}^T}{|\mathbf{w}|^2}. \quad (26)$$

In this way, $\mathbf{w}_{\text{observed}}$ becomes an eigenvector of C_{11}^* and is therefore in the span as desired.

Provided that α is sufficiently small relative to the trace of C_{11} , the spatial correlations in the resulting simulation will not be adversely affected. If C_{11} is adjusted C_{21} must be adjusted

correspondingly to maintain the consistency between C_{11} and C_{21} . Given C_{11}^* we wish to find

some matrix C_{21}^* such that $C_{11}^* = F C_{21}^*$. This linear system can be solved by singular value

decomposition

$$F = X \Lambda Y^T. \quad (27)$$

Define $F^+ = Y\Lambda^+X^T$, where $\Lambda^+ = \text{diag}(\delta_1^{-1}, \delta_2^{-1}, \delta_3^{-1}, \dots, \delta_k^{-1}, 0, \dots, 0)$ is the diagonal matrix of reciprocals of singular values with small values set to zero. This is a generalized inverse for non square matrices having the property that $FF^+F = F$, since the columns of Y and X are mutually orthogonal (Press et al 1994). With this construction we have

$$C_{21}^* = F^+C_{11}^*. \quad (28)$$

The primary limitation is that in general F may be large and calculating the SVD may be prohibitive in practice. An alternative could include application of iterative techniques to solve for C_{21}^* . Note that this system need only be solved once so an iterative procedure here does not severely impact the overall efficiency given that multiple realizations are to be simulated. The degree to which C_{11} and C_{21} must be consistent has not been fully investigated. In most cases, conditioning on the projection as opposed to the actual data is probably an adequate solution.

Comparison of Algorithms

We conducted a set of empirical comparisons between the sequential Gaussian method and the FFT method for simulating multivariate Gaussian random fields. We compared the accuracy of reproduction of the specified mean of the simulation, sample variance, and sample covariance function. Mean and variance were compared based on the average bias, average absolute bias and mean squared error (MSE). The reproduction of the covariance functions were compared graphically. We also compared the time required to produce 1000 realizations from each algorithm with 20, 40, 60 and 80 conditioning points, and using small and large search neighborhoods for the sequential algorithm.

RESULTS

Summary statistics for the mean and variance of 1000 simulated realizations are reported in Tables 1 and 2. As the size of the realization increases from a square of size 2 zones of influence to 32 zones of influence, the bias in the mean and variance decreases for both algorithms. The bias, absolute bias and mean squared error of the variance are similar for the frequency domain and sequential algorithms. In contrast, the bias absolute bias and mean squared error for the stationary mean parameter are lower for the frequency domain method than for the sequential method with the absolute bias for the sequential method 7 orders of magnitude higher than that for the frequency domain method. In general the absolute bias for the frequency domain method is near the magnitude of roundoff error for the algorithm implementation, while the sequential method allows the realization mean to *wander* substantially more. This tendency to wander is particularly high for smaller simulations such as the grids with less than 8 zones of influence.

Sample covariance functions from 10 realizations are plotted in figure 2. Both algorithms reproduced the covariance function similarly. The true covariance function was an exponential model with no nugget effect, range of influence one unit and variance parameter one.

Aside from the difference in reproduction of the mean parameter, algorithm speed is the only criteria in which marked differences are apparent. However this difference must be qualified as the differences in speed are related to the number of conditioning points for the frequency domain method and how the search parameters are set up for the sequential algorithm. In spite of these intricacies, there appear to be some conclusions which can be drawn. For all grid sizes investigated, 16 by 16 up to 256 by 256, the frequency domain method was faster regardless of the search neighborhood parameters. The gain in speed was less when the search

radius was restricted for the sequential algorithm. The speed of the sequential algorithm is independent of the number of conditioning points for a fixed search neighborhood and system size.

When the restricted search neighborhood was used for the sequential algorithm, the FFT method was 17 to 27 times faster with 20 conditioning points, and approximately 4 times faster with 80 conditioning points. When the larger search neighborhood was utilized, the speed differentials were much more dramatic. When 80 conditioning points were used the speed gain was approximately 80 times. Gains in speed increased as the number of samples decreased.

DISCUSSION

In general, both algorithms performed similarly with regard to reproduction of the theoretical model although the sequential algorithm tended to allow the realization mean to wander more than the FFT method. We view this as somewhat of a concern as the variation in the process mean should be based on the sampling distribution of the estimated mean as opposed to algorithm fluctuations. Using the FFT method, the process mean for each realization should be a random draw from the sampling distribution based on the data. In situations where sample sizes are low to moderate (20 to 80) as is often the case with environmental studies, the frequency domain methods are faster in a practical sense. For example to consider a 128 by 128 grid with 40 conditioning points, the frequency domain analysis took less than 30 minutes for 1000 realizations, while the sequential analysis took a minimum of 4 hours of processor time on a moderately powerful desktop personal computer. The frequency domain technique reduces a large class of Monte Carlo problems from an overnight computer run to one that runs in a few minutes. This may be of practical importance in many situations.

It is clear that both algorithms have a useful place in analysis of environmental data. For problems with several hundred data locations, the sequential algorithm or the iterative FFT methods may be preferable. However, when several tens to one hundred data locations are available, the frequency domain method appears to be the method of choice. An additional advantage of the frequency domain method over the sequential method is that there are no parameters to choose once the correlation model has been estimated. The simulation algorithm can be determined directly from parameters estimated from the data. Conversely, for the sequential algorithm there are several parameters related to how the search strategy is to be implemented which must be chosen. This is the same conundrum which is faced when kriging algorithms are used for smoothing and prediction. One is faced with a series of choices the implications of which may be poorly understood. Justification of choices in algorithm fine tuning and assumptions underlying the choices make good fodder for opposing sides to pick apart. When the frequency domain method is employed, there are 2 basic assumptions, that the distribution can be transformed to a normal distribution, and that second order stationarity of the transformed data is appropriate. There are no questions regarding whether data should have more weight than simulated nodes, what shape of search neighborhood should be used to attain the correct anisotropy, how large the search neighborhood should be etc. This reduction in arbitrary decisions may add significantly to the level of scientific objectivity and credibility of the investigations. As with conventional statistical methods, one data set will give rise to one inferential conclusion regardless of who conducts the analysis. We see this as a significant advantage for applications in risk assessment and other environmental applications where litigation is likely.

ACKNOWLEDGMENTS

We wish to thank the United States Department of Interior Bureau of Land Management, and Office of Surface Mining Reclamation and Enforcement for funding through a Wyoming Abandoned Coal Mine Lands Research Program grant administered by the University of Wyoming. We also thank the United States Department of Energy for grants through Montana State University.

REFERENCES

- Borgman, L.E. 1982. Techniques for computer simulation of ocean waves. *Advanced Topics in Ocean Physics*. Italian Physical Society, Bologna, Italy.
- Borgman L. E., Taheri, M. and Hagan, R. 1984. Three-dimensional, frequency-domain simulations of geologic variables. In *Geostatistics for Natural Resources Characterization, Part 1*, G. Verly, M. David, A. G. Journel and A. Marechal, eds. Reidel, Dordrecht.
- Borgman, L. E. and R. C. Faucette. 1993a. Basic mathematics and statistical theory for finite Fourier coefficients of Gaussian vector functions. In: *Computational Stochastic Mechanics, Chap 1*. Computational Mechanics Publications, London.
- Borgman, L. E. and R. C. Faucette. 1993b. Multidimensional simulation of Gaussian vector random functions in frequency domain. In: *Computational Stochastic Mechanics, Chap 2*. Computational Mechanics Publications, London.
- Borgman, L.E. C. D. Miller, S.R. Signorini, and R.C. Faucette. 1994. Stochastic interpolation as a means to estimate oceanic fields. *Atmosphere Ocean*. V. 32., n. 2. p. 395-419.
- Davis, M.W., 1987. Production of Conditional Simulations via the LU decomposition of the

- covariance matrix. *Math. Geol.* V. 19, n. 2, p. 91-98.
- Deutsch, C. V. and A. G. Journel. 1992. *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, New York.
- Easley, D.H., L.E. Borgman and D. Weber, 1991. Monitoring well placement using conditional simulation of hydraulic head, *Math. Geol.* V. 23, n. 8, p 1059-1081.
- Journel, A.G., and C.J., Juijbregts, 1978. *Mining Geostatistics*, Academic Press, New York.
- Press, W., B. Flannery, S. Teulkolsky, and W. Vetterling. 1986. *Numerical Recipes*. Cambridge University Press, New York
- Taheri, G.B., 1980. Data retrieval and multidimensional simulation of mineral resources. PhD. Dissertation, University of Wyoming, Department of Statistics, Laramie WY.
- Yao, T. 1998. Conditional spectral simulation with phase identification. *Mathematical Geology*, Vol. 30, No 3. p 285-308.

Table 1. Bias, absolute bias, and mean squared error for sample mean and variance from conditionally simulated random fields using the frequency domain algorithm. The grid spacing was fixed at 1/8 range of correlation and the grid size was varied from 2 ranges of influence through 32 ranges of influence in the horizontal and vertical directions.

		Statistics for The Mean			Statistics for the Variance		
Length of Side		Absolute			Absolute		
of Square Grid	Bias	Bias	MSE	Bias	Bias	MSE	
$2h_0$	-6.98×10^{-11}	3.30×10^{-09}	1.75×10^{-17}	-0.0694	0.1805	0.0494	
$4 h_0$	1.12×10^{-10}	2.17×10^{-09}	7.85×10^{-18}	-0.0311	0.1139	0.0204	
$8 h_0$	2.86×10^{-11}	1.34×10^{-09}	2.98×10^{-18}	-0.0011	0.0632	0.0062	
$16 h_0$	-2.15×10^{-11}	7.86×10^{-10}	9.73×10^{-19}	0.0003	0.0311	0.0015	
$32 h_0$	-2.85×10^{-11}	4.31×10^{-10}	2.86×10^{-19}	0.0001	0.0164	0.0004	

Table 2. Bias, absolute bias, and mean squared error for sample mean and variance from conditionally simulated random fields using the sequential simulation algorithm. The grid spacing was fixed at $1/8$ range of correlation and the grid size was varied from 2 ranges of influence through 32 ranges of influence in the horizontal and vertical directions.

Length of Side of Square Grid	Statistics for The Mean			Statistics for the Variance		
	Bias	Absolute		Bias	Absolute	
		Bias	MSE		Bias	MSE
$2 h_0$	0.0028	0.2633	0.1096	-0.1297	0.2034	0.0566
$4 h_0$	0.0084	0.1479	0.0351	-0.0363	0.1094	0.0179
$8 h_0$	-0.0041	0.0779	0.0097	-0.0143	0.0560	0.0051
$16 h_0$	0.0025	0.0398	0.0025	-0.0076	0.0280	0.0012
$32 h_0$	0.0007	0.0224	0.0007	-0.0010	0.0143	0.0003

Table 3. Simulation times for the frequency domain and sequential algorithms measured in seconds per 1000 simulations. Algorithm speed is independent of number of conditioning points for the sequential algorithm although the search neighborhood parameters do have significant influence on algorithm speed.

Grid Size	Frequency Domain Algorithm				Sequential Algorithm	
	20 points	40 points	60 points	80 points	Small	Large
					Nbhd ^a	Nbhd ^b
16 by 16	10	21	52.	55	270	4390
32 by 32	51	63	167	236	1,004	19,670
64 by 64	223	343	629	944	4,010	79,420
128 by 128	894	1649	2541	3713	16,040	320,220
256 by 256	3,688	6,794	10,167	14,860	63,940	1,306,000

^aThe smaller search neighborhood consisted of limiting the system size to 10 simulated nodes and 10 data values.

^bThe large search neighborhood consisted of a maximum of 64 simulated nodes and 64 data locations.

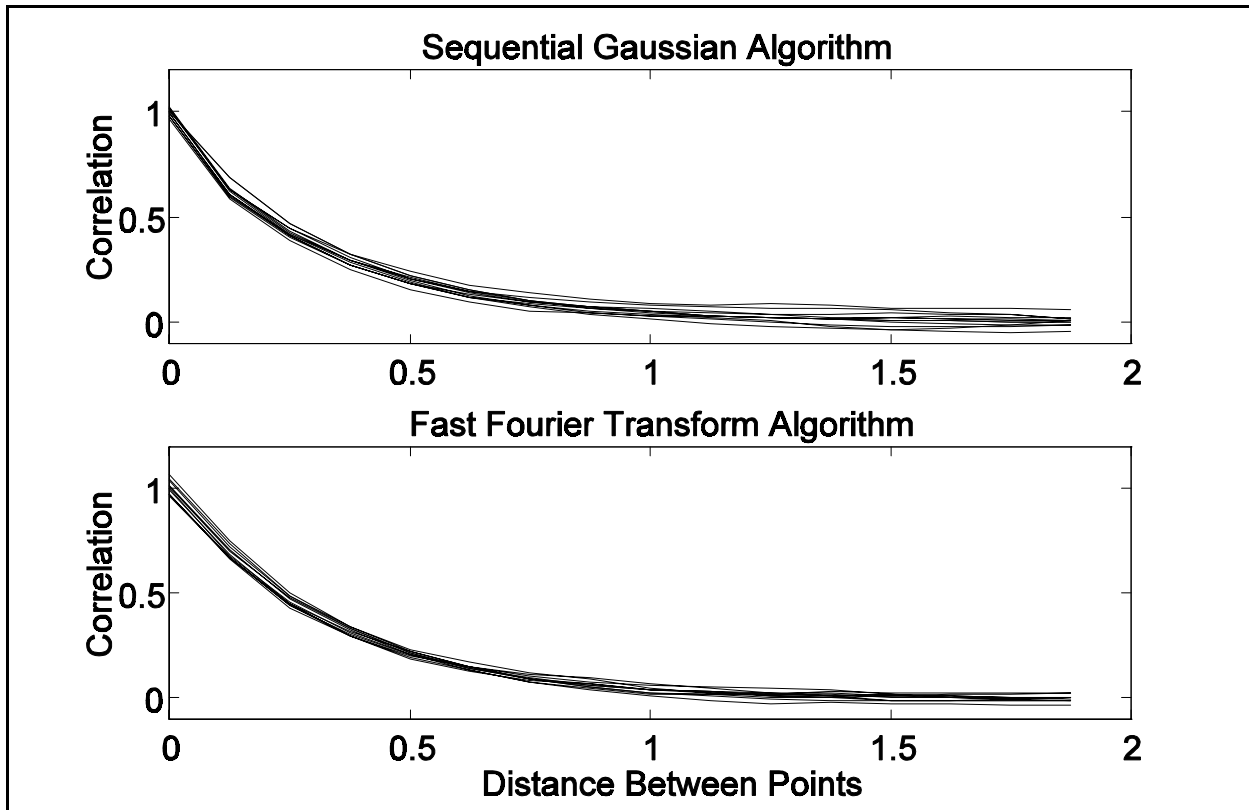


Figure 2. Plot of estimated auto-correlation functions in the North South Direction from 10 simulated realizations generated by the sequential Gaussian algorithm (top panel) and the FFT algorithm (bottom panel). The simulation grid contained 128 rows and columns with grid spacing $1/8$ the range of influence. The resulting grid was 16 ranges of influence on a side.