Impact of Data Spacing on Variogram Uncertainty

Hadi Derakhshan and Oy Leuangthong

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

The variogram model used in estimation and/or simulation is unquestionably important. Uncertainty in the variogram must be understood and accounted for in fitting. This may not be a problem with many data; however, geostatisticians are commonly faced with sparse data and significant uncertainty in the variogram model. This paper examines and quantifies the impact of data spacing on variogram uncertainty. A synthetic reference model is considered with a known variogram model. Using this reference model, different sample spacings and data configurations are considered for variogram calculation. For each case, a variogram model is fit and the variance of the variogram model and experimental points is quantified. This provides insight into the variability that may be expected in the fitted model as a function of the data spacing. These numerical results are compared to theoretical models of uncertainty.

Introduction

While the emergence of multiple point statistics (MPS) has spawned much research into multipoint geostatistics, a large faction of practical and theoretical geostatistics remains deeply reliant on the variogram. Unlike MPS, the variogram is a two-point statistic that spatially relates two random variables (RV), $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$:

$$2\gamma(\mathbf{h}) = E\left\{[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})]^2\right\}$$

where **u** and **h** are location and lag vectors, respectively, in domain A. Matheron (1965) first proposed a method-of-moments approach to approximate the variogram:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[z_i(\mathbf{u}) - z_i(\mathbf{u}+\mathbf{h})\right]^2$$

where N(**h**) is the number of pairs of data separated by a vector **h**. This numerical approximation laid the foundation for most theoretical and practical development in the area of variogram modeling and uncertainty.

Given its importance in geostatistical methods such as change of support, kriging and simulation, it is not surprising that the issue of variogram uncertainty and fitting has been extensively covered in the literature. Davis and Borgman (1979) developed the characteristic function of the variogram estimator, $\hat{\gamma}(\mathbf{h})$, for an equally-spaced, one-dimensional, stationary Gaussian random function (RF) model. They tabulated the sample distribution of the variogram estimator, using a Finite Fourier Transform (FFT) inversion. In 1982, Davis and Borgman further proved that the distribution of the sample variogram is indeed asymptotic:

$$L\left\{\frac{\hat{\gamma}(\mathbf{h})-\gamma(\mathbf{h})}{\sigma[\gamma(\mathbf{h})]}\right\} \to N(0,1) \text{ as } N(\mathbf{h}) \to \infty$$

Many authors have focused on the derivation of the variance/covariance matrix of the experimental variogram, mainly with the purpose to determine an optimum fit for the variogram. David (1977) proposed the use of an ordinary least squares approach to minimize

$$\sum_{i=1}^{nh}\left[\hat{\gamma}(\mathbf{h}_i)-\gamma(\mathbf{h}_i)\right]^2$$

where *nh* is the number of lags. Cressie (1985) later approximated the variance of the variogram estimates for a Gaussian variable as

$$\text{Var}[\hat{\gamma}(\mathbf{h})] \simeq \frac{2[2\gamma(\mathbf{h})]^2}{N(\mathbf{h})} \qquad (1)$$

These were then used for variogram fitting using a weighted least squares (WLS) approach, where the weights account for the numbers of pairs within each class. It can be shown that the variogram estimator, for a Gaussian variable, is a linear combination of independent χ-square random variables, each with one degree of freedom (Cressie, 1993):

$$2\hat{\gamma}(\mathbf{h}) = \sum_{i=1}^{n}\lambda_i(\mathbf{h})\chi^2_{1,i}$$

Cressie goes on to show the use of this result for the robust estimation of the variogram.

Ortiz and Deutsch (2000) developed an analytical expression for the pointwise variogram uncertainty. They calculated the uncertainty in the variogram by assuming a known variogram model. They showed that the uncertainty in the variogram is the average covariance between pairs of pairs used to calculate the variogram for the particular lag:

$$\sigma^2_{2\hat{\gamma}(\mathbf{h})} = \frac{1}{N^2(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\sum_{j=1}^{N(\mathbf{h})}C_{ij}(\mathbf{h})$$

where

$$C_{ij}(\mathbf{h}) = Cov\left\{\left[Z(\mathbf{u}_i)-Z(\mathbf{u}_i+\mathbf{h})\right]^2, \left[Z(\mathbf{u}_j)-Z(\mathbf{u}_j+\mathbf{h})\right]^2\right\}$$

Pardo-Igúzquiza and Dowd (2001) examined the variance-covariance matrix of the experimental variogram, $\left[Cov\{\hat{\gamma}(\mathbf{h}),\hat{\gamma}(\mathbf{h}')\}\right]$. Similar to Ortiz and Deutsch's (2000) approach, this required examination of a fourth order statistic; however, the expression developed by Pardo-Igúzquiza and Dowd accounts for the joint uncertainty of the variogram between two different lags.

Marchant and Lark (2004) conducted simulation experiments to estimate variogram uncertainty for two simulation field sizes and three different sampling schemes. Integral to their approach was use the use of a generalized least squares (GLS) approach to variogram fitting. Based on these experiments, they concluded that Pardo-Igúzquiza and Dowd's approach (2001) provided a good estimate of variogram uncertainty due to ergodic errors.

The purpose of this study is to quantify the relationship between the variogram uncertainty and the data spacing for the special case of regular sample spacing. A synthetic reference model is constructed, and several data spacings and configurations are considered for variogram calculation and modeling. Relating variogram uncertainty to data spacing is illustrated by examining the variance of the variogram as a function of distance. Depending on the data spacing, one can generate a similar diagnostic chart to examine whether the fitted model falls within the expected variability of the variogram given the available data spacing.

Background

One can relate each point on a variogram plot to an **h**-scatterplot, which shows all possible pairs of data values whose locations are separated by a certain distance vector **h**. Journel (1989) described the calculation of the variogram from this **h**-scatterplot as calculating the moment of inertia about the 45° line (see Figure 1).

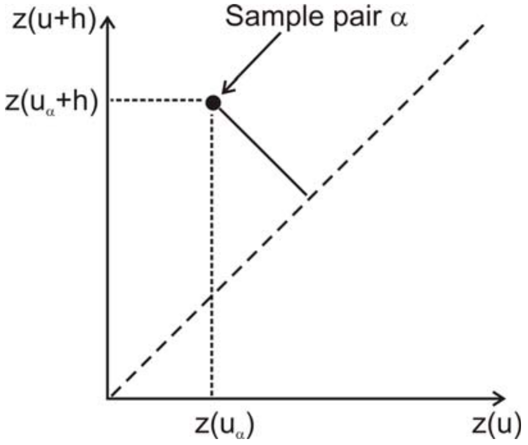

Figure 1: Moment of inertia interpretation of the variogram based on an h-scatterplot. (Redrawn from Goovaerts, 1997).

Based on the distribution of the cloud of points on an **h**-scatterplot, we can tell how similar the data values are over a certain distance in a particular direction (as defined by the lag vector **h**). If the data values at locations separated by **h** are similar, the pairs will plot close to a 45° line. We naturally expect that this cloud of points will show little dispersion at small lag distances, but as **h** increases, this cloud of paired values is expected to increase in dispersion. This notion of dissimilarity (or dispersion) is neatly captured by the variogram.

The number of pairs available for computing the variogram depends on the lag distance. For regularly spaced samples, as the lag separation gets larger there are fewer points, so the method-of-moments approximation for the variogram is less precise at larger lag distances. If there are *n* observed data, then there are $n(n-1)/2$ unique pairs of observations taken over all possible lag distances. Thus, even a data set of moderate size generates a large number of pairs. For example, if 500 samples are available, there are 124,750 pairs of data if we considered all lags simultaneously. Figure 2 shows a few different lag distances in the case of regular spaced data for calculating the experimental variogram. We can see that depending on the direction, the lag spacing considered, and the size of the regular grid, the number of pairs used for calculating the variogram can be quite different.

In practice, data are rarely exactly regularly spaced. Sampling campaigns may target nominal drillhole/well spacing; however, certain regions of the deposit/reservoir are inevitably more densely drilled as they provide more information about the available resource. As such, real data are irregularly spaced and the paired information used in calculating the experimental variogram are based on *approximate* lag separation distances. Bandwidths about the desired direction, along with angle and lag tolerances are often considered (Deutsch and Journel, 1998).

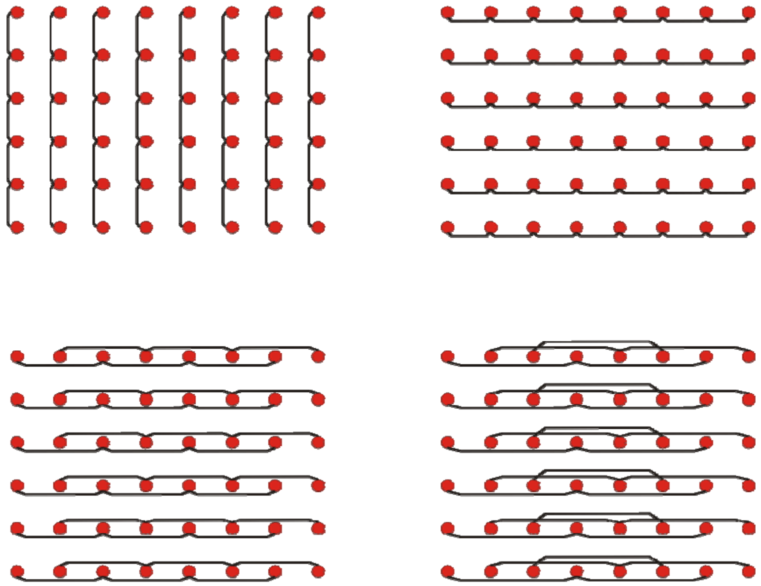

Figure 2: Different Lag Distances in the Case of Regular Spaced Data: h=1 taken vertically yields 40 pairs (top left); h=1 taken horizontally results in 42 pairs (top right); h=2 in the horizontal direction will give 36 pairs (bottom left); and h=3 in the horizontal direction results in 30 pairs (bottom right).

Even after the variogram is numerically calculated, we must still fit the experimental points with a positive semi-definite variogram model. This model is then used in subsequent estimation and/or simulation. Theoretically, we are not constrained to consider any set of models so long as positive semi-definiteness of the resulting model is ensured. Practically, this can be quite prohibitive given the challenges associated to validating that this positive semi-definiteness condition is guaranteed for all directions and all distances. As a result, there are a set of theoretically validated models that are widely adopted including the nugget, spherical, exponential and Gaussian models. These can be linearly combined in an infinite number of ways to fit most experimental variograms. Gringarten and Deutsch (2001) provide an extensive discussion on variogram interpretation and some guidelines on variogram modeling.

Of course, the uncertainty in calculating an experimental variogram is carried forward and somehow resolved by the user when the experimental points are fit with a licit model. Specifically there are some key components of the fit that are important:

- Although the value of the variogram for **h** = 0 is strictly zero, short scale variability may cause sample values separated by extremely small distances (lag) to be quite dissimilar. This result in an apparent vertical intercept on the variogram plot that is often referred to as the nugget effect.

- For a stationary random function, the limit of 'dissimilarity' or the variogram value at which the variogram points appear to converge to at large lag distances is referred to as the sill. We can also interpret the sill as the value at which paired data are no longer correlated to each other, or $C(\mathbf{h})=0$ where $C(\mathbf{h})$ is the covariance of pairs of data separated by **h**. The well established relationship between the variogram, covariance and variance, $\gamma(\mathbf{h}) = C(0) - C(\mathbf{h})$, where C(**0**) represents the variance, demonstrates that the sill of the variogram is equivalent to the variance of the data:

$$\gamma(\infty) = C(0) = \sigma^2$$

- The range is the lag distance at or near which the variogram reaches the sill; beyond that distance the corresponding correlation coefficient is zero.

The next section describes the problem setting to evaluate the uncertainty in the variogram as a function of data spacing. A very specific scope is considered for this study, and a small example is provided for additional insight into this relationship and comparisons with earlier approximations from previous authors.

Problem Setting

Consider a 2D domain that is discretized into an *n* x *n* grid, for which samples are available at a regular spacing of *m* x *m* units. While regular sample spacings are not common in practice, considering this very particular case permits us to examine several interesting issues related to experimental variogram uncertainty. The calculation of variogram uncertainty is dependent on the data configuration and the available number of data pairs found for a specific lag. In fact, for this special case of regular sample spacing, we can quantify exactly (1) the number of data configurations given specific sample spacing, and (2) the available number of data pairs that is often used to gauge the reliability of a specific variogram value.

Data configuration is a function of the sample spacing and the field size. For regular data spacing within a square grid, the number of possible configurations can be generalized:

$$No.\,configurations = \begin{cases} m^2 & for \quad 1 \le m \le \left\lfloor \dfrac{n}{2} \right\rfloor \\ (n-m)^2 & for \quad \left\lfloor \dfrac{n}{2} \right\rfloor < m \le n-1 \end{cases}$$

where *m* is the data spacing, *n* is the size of the discretized grid, and $\lfloor x \rfloor$ is the floor function and is equal to the maximum integer number which is less than or equal to *x*. For instance, for 2x2 data spacing in a 1024 x 1024 grid, there are four possible configurations in which the samples could have been obtained; for a 3x3 spacing in the same grid, nine sample configurations are possible (see Figure 3). The extension of the four possible 2x2 spacing configurations to the 1024x1024 grid considered in this example is also shown in Figure 4. Based on the above equation, Figure 5 shows that the number of possible configurations increases and decreases in a quadratically symmetric fashion about the mid-grid data spacing.

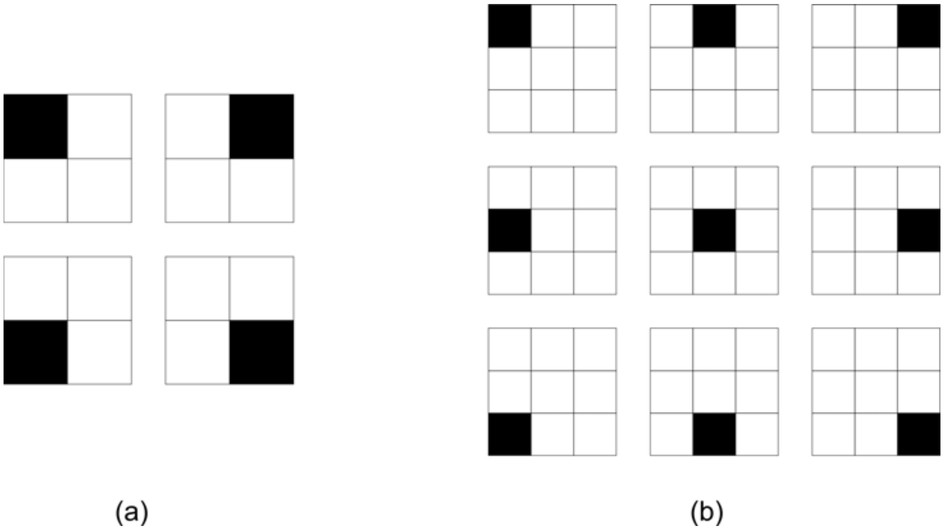

Figure 3: Different sample configurations for same data spacing: (a) 2 x 2 spacing, and (b) 3x3 spacing.

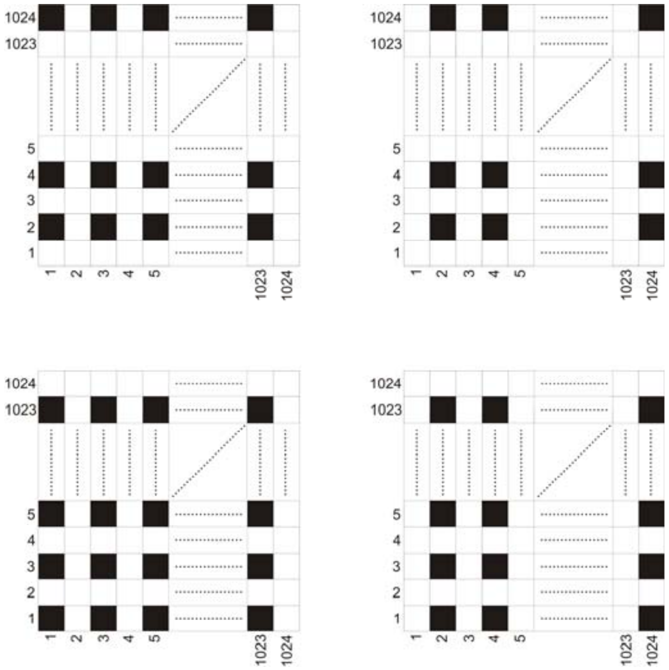

Figure 4: Extension of 2x2 data configuration on a 1024x1024 grid.

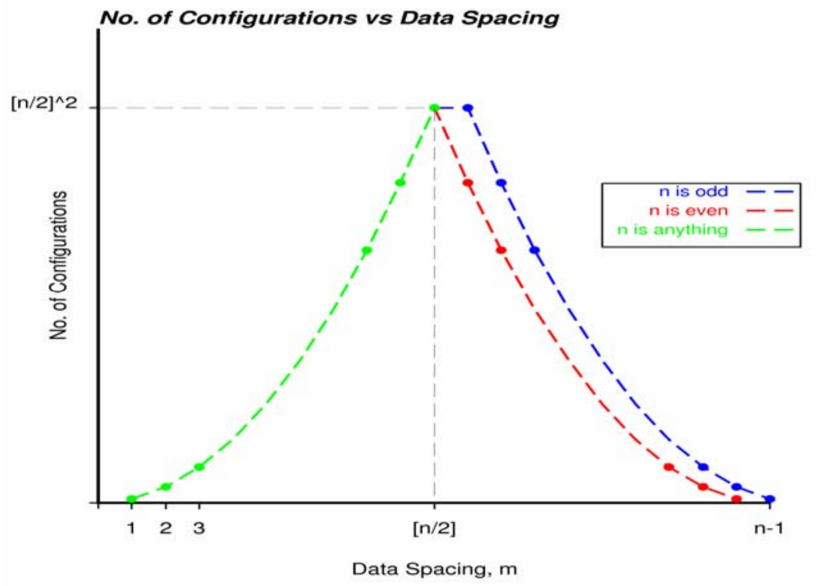

Figure 5: Number of configurations in the case of *m*x*m* data spacing, in an *n*x*n* field.

In this reasonably well-controlled example, the number of pairs of data can also be determined. This information is often used in practice as an indicator of the reliability of an experimental variogram value. In fact, the number of pairs found at a lag vector **h** in this 2D case is a function of the following seven parameters:

- n_x and n_y which represent the size of the discretized field in the *x* direction and *y* direction, respectively;
- m_x and m_y corresponding to the sample spacing in the *x* direction and *y* direction, respectively;
- **h** which is the lag vector, and in this case of regular sample spacings, **h** is an integer multiple of the sample spacing in a specific direction, i.e. $\mathbf{h}_x = m_x$, $2m_x$, and so on.
- *i* and *j* which correspond to indices associated to the first sample closest to the origin of the grid (based on GSLIB grid definition). A simple schematic illustrating the specification of the indices is shown for the 2x2 data spacing scenario:

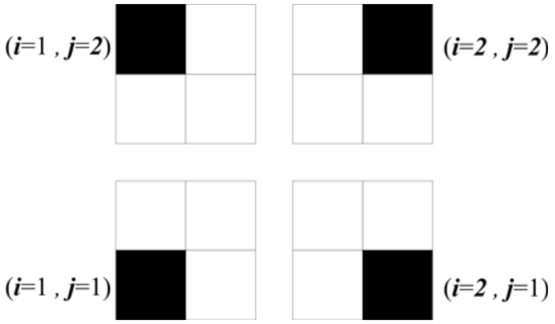

Figure 6: Indices *i*,*j* denote the configuration of samples based on a certain sample spacing; shown here for 2 x 2 data spacing.

In general we can obtain the number of pairs found at a lag vector **h** in the *x* and *y* directions by:

$$\begin{cases} N_x\left(\mathbf{h};n_x,n_y,m_x,m_y,i,j\right)=\left\lfloor \dfrac{n_x+m_x-\mathbf{h}-i}{m_x}\right\rfloor\left\lfloor \dfrac{n_y+m_y-j}{m_y}\right\rfloor \\ N_y\left(\mathbf{h};n_x,n_y,m_x,m_y,i,j\right)=\left\lfloor \dfrac{n_x+m_x-i}{m_x}\right\rfloor\left\lfloor \dfrac{n_y+m_y-\mathbf{h}-j}{m_y}\right\rfloor \end{cases}$$

If we assumed an omnidirectional variogram at a spacing of **h**, then the number of pairs in both directions can simply be added together:

$$\begin{aligned} N\left(\mathbf{h};n_x,n_y,m_x,m_y,i,j\right)=&\left\lfloor \frac{n_x+m_x-\mathbf{h}-i}{m_x}\right\rfloor\left\lfloor \frac{n_y+m_y-j}{m_y}\right\rfloor \\ &+\left\lfloor \frac{n_x+m_x-i}{m_x}\right\rfloor\left\lfloor \frac{n_y+m_y-\mathbf{h}-j}{m_y}\right\rfloor \end{aligned} \tag{2}$$

In the special case where n_x and n_y can be neatly divided by m_x and m_y respectively, the above formula is further simplified and *i* and *j* can be removed from the formula:

$$N_x\left(\mathbf{h};n_x,n_y,m_x,m_y\right)=\left(\frac{n_x-\mathbf{h}}{m_x}\right)\left(\frac{n_y}{m_y}\right)$$

$$N_y\left(\mathbf{h};n_x,n_y,m_x,m_y\right)=\left(\frac{n_x}{m_x}\right)\left(\frac{n_y-\mathbf{h}}{m_y}\right)$$

$$N\left(\mathbf{h};n_x,n_y,m_x,m_y\right)=\frac{1}{m_x m_y}\left[n_y\left(n_x-\mathbf{h}\right)+n_x\left(n_y-\mathbf{h}\right)\right] \tag{3}$$

and if $n_x = n_y = n$ and $m_x = m_y = m$, then

$$N\left(\mathbf{h};n,m\right)=\frac{2}{m^2}n\left(n-\mathbf{h}\right) \tag{4}$$

The extension to a 3D case, where the grid is discretized into $n_x \times n_y \times n_z$ locations, with a regular data spacing of $m_x \times m_y \times m_z$, is straightforward. The number of pairs for an omnidirectional is given as:

$$
\begin{aligned}
N\left(\mathbf{h};n_x,n_y,n_z,m_x,m_y,m_z,i,j,k\right)=&\left\lfloor\frac{n_x+m_x-\mathbf{h}-i}{m_x}\right\rfloor\times\left\lfloor\frac{n_y+m_y-j}{m_y}\right\rfloor\times\left\lfloor\frac{n_z+m_z-k}{m_z}\right\rfloor+\\
&\left\lfloor\frac{n_x+m_x-i}{m_x}\right\rfloor\times\left\lfloor\frac{n_y+m_y-\mathbf{h}-j}{m_y}\right\rfloor\times\left\lfloor\frac{n_z+m_z-k}{m_z}\right\rfloor+\\
&\left\lfloor\frac{n_x+m_x-i}{m_x}\right\rfloor\times\left\lfloor\frac{n_y+m_y-j}{m_y}\right\rfloor\times\left\lfloor\frac{n_z+m_z-\mathbf{h}-k}{m_z}\right\rfloor
\end{aligned}
$$

Equation (3) becomes:

$$
N\left(\mathbf{h};n_x,n_y,n_z,m_x,m_y,m_z\right)=\frac{1}{m_x m_y m_z}\left[n_y n_z\left(n_x-\mathbf{h}\right)+n_x n_z\left(n_y-\mathbf{h}\right)+n_x n_y\left(n_z-\mathbf{h}\right)\right]
$$

and Equation (4) becomes

$$
N\left(\mathbf{h};n,m\right)=\frac{3}{m^3}n^2\left(n-\mathbf{h}\right)
$$

Further, the above formula can be generalized for the *d*-dimensions:

$$
N\left(\mathbf{h};n,m\right)=\frac{d}{m^d}n^{d-1}\left(n-\mathbf{h}\right)
$$

The Variance of the Variogram in 2-D case

Given an *m* x *m* data spacing, there are m^2 possible configurations and hence m^2 possible values for the experimental variogram. Thus the variance of these m^2 values for variograms at each **h** can be calculated as

$$
\sigma^2_{\hat{\gamma}(\mathbf{h})}\left(\mathbf{h},m\right)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left[\hat{\gamma}_{i,j}\left(\mathbf{h}\right)-\overline{\hat{\gamma}\left(\mathbf{h}\right)}\right]^2}{m^2}
$$

where

$$
\hat{\gamma}_{i,j}\left(\mathbf{h}\right)=\frac{1}{N\left(\mathbf{h};n_x,n_y,m,i,j\right)}\sum_{k=1}^{N\left(\mathbf{h};n_x,n_y,m,i,j\right)}\left[z_{i,j}\left(\mathbf{u}_k\right)-z_{i,j}\left(\mathbf{u}_k+\mathbf{h}\right)\right]^2
$$

and

$$
\overline{\hat{\gamma}\left(\mathbf{h}\right)}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\hat{\gamma}_{i,j}\left(\mathbf{h}\right)}{m^2}
$$

Note that $\sigma^2_{\hat{\gamma}(\mathbf{h})}(\mathbf{h},m)$ is the variance of the experimental variogram as a function of lag distance (**h)** and data spacing (*m*), $\hat{\gamma}_{i,j}(\mathbf{h})$ is the experimental variogram value at lag distance, **h**, for the configuration of (*i* , *j*) generated by *m* x *m* data spacing. $\overline{\hat{\gamma}(\mathbf{h})}$ is the average of the experimental variogram values at lag vector of **h** over *m*^{2} possible values for experimental variogram.

Recall that Cressie (1985) approximated the variance of the variogram for a Gaussian variable (see Equation (1)). Given the special case of a regular field where $n_x = n_y = n$, and *n* can be neatly divided by the data spacing, *m*, Cressie's model can be simplified to

$$\sigma^2_{\hat{\gamma}(\mathbf{h})}(\mathbf{h},m) \simeq \frac{\left[2m\gamma(\mathbf{h})\right]^2}{n(n-\mathbf{h})} \tag{5}$$

The following example compares Cressie's approximation to the experimental results based on a square grid with regular sample spacing.

Example

A synthetic 2D Gaussian random field (Figure 7) is generated via an unconditional simulation for a 1024 x 1024 grid with the following reference variogram:

$$\gamma(\mathbf{h}) = 0.05 + 0.95Sph_{a=64}(\mathbf{h})$$

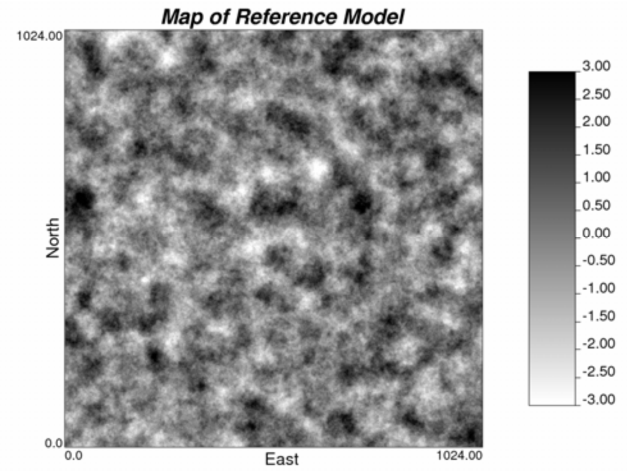

Figure 7: Map of reference model generated by SGSIM.

Using this reference model, we can then sample at various data spacings and calculate the corresponding experimental variogram. The specific procedure to examine these scenarios is summarized below:

1. Sample the reference model at specific data spacings, such as 2x2, 3x3, and so on.
2. For each possible configuration given the data spacing:
    a. Calculate the experimental variogram for each possible configuration.

b. Fit the experimental variogram using a licit model. In cases where the number of possible configurations is reasonably small, a manual fitting can be performed; however, as the data spacing becomes larger and the number of possible configurations becomes prohibitively large for manual fitting, a (semi) automatic variogram fitting algorithm could also be considered. The results shown in this study used the VARFIT program for this task (Larrondo, Neufeld and Deutsch, 2003; Neufeld and Deutsch, 2004).

3. Consider all the resultant variogram models, calculate and plot the variance of the variogram, $\sigma^2[\gamma(\mathbf{h})]$, as a function of lag distance, **h**.

All the tasks above were performed using a combination of GSLIB (Deutsch and Journel, 1998) and GSLIB-compatible programs.

Figure 8 shows the experimental variogram plots for 2x2, 4x4, 8x8, 16x16, 32x32 and 64x64 data spacings up to range of 512. It is obvious that for 2x2 data spacing the lag distance starts from 2, for 4x4 from 4, and so on. Therefore for the 64x64 spacing, there are no experimental points for a lag distance less than 64m. The solid line in each plot represents the reference variogram model. Clearly, variograms based on the 64x64 data spacing are the most uncertain; this is not surprising given that the variogram range coincides with the data spacing.

The distribution of experimental variogram values at **h**=192 in the case of 64x64 data spacing is shown in Figure 9. Although, the specific case of 64x64 has already been deemed to be the least interesting given the variogram range, the distribution of the experimental variogram is interesting to consider. We see that it is an approximately normal distribution, and the reference value of 1.0 based on the variogram model is approximately equal to the upper quartile. Further discussions of the variance of the variogram are precisely based on the variance obtained from lag histograms of the experimental variogram such as this one.

Figure 10 shows Cressie's model for six different data spacings (2x2, 4x4, 8x8, 16x16 , 32x32 and 64x64) based on the same reference variogram for this 1024x1024 field. We see that at a constant **h,** as data spacing, *m*, increases, the variance of the variogram also increases. At constant data spacing, *m*, as **h** increases the variance increases. This increase is more pronounced below the reference range (**h**=64), and then dramatically flattens beyond the range. Similar to the experimental case, for a particular mxm data spacing, there is no value given for $\mathbf{h}\leq m$; this is not surprising, given the formula that we used for the number of pairs **h** apart in the case of regular spaced data (see Equations 4 and 5). Interestingly, the difference between each two curves that can be simply quantified as:

$$\log\left[\sigma^2_{\hat{\gamma}(\mathbf{h})}\left(\mathbf{h},m_2\right)\right]-\log\left[\sigma^2_{\hat{\gamma}(\mathbf{h})}\left(\mathbf{h},m_1\right)\right]=2\times\log\frac{m_2}{m_1}$$

Particular to this example, the ratio of $\frac{m_2}{m_1}$ is constant and equal to 2 (assuming that we consider the spacing sequentially as 2, 4, 8, 16, 32 and 64), so the above equation reduces to

$$\log\left[\sigma^2_{\hat{\gamma}(\mathbf{h})}\left(\mathbf{h},m_2\right)\right]-\log\left[\sigma^2_{\hat{\gamma}(\mathbf{h})}\left(\mathbf{h},m_1\right)\right]=2\times\log 2$$

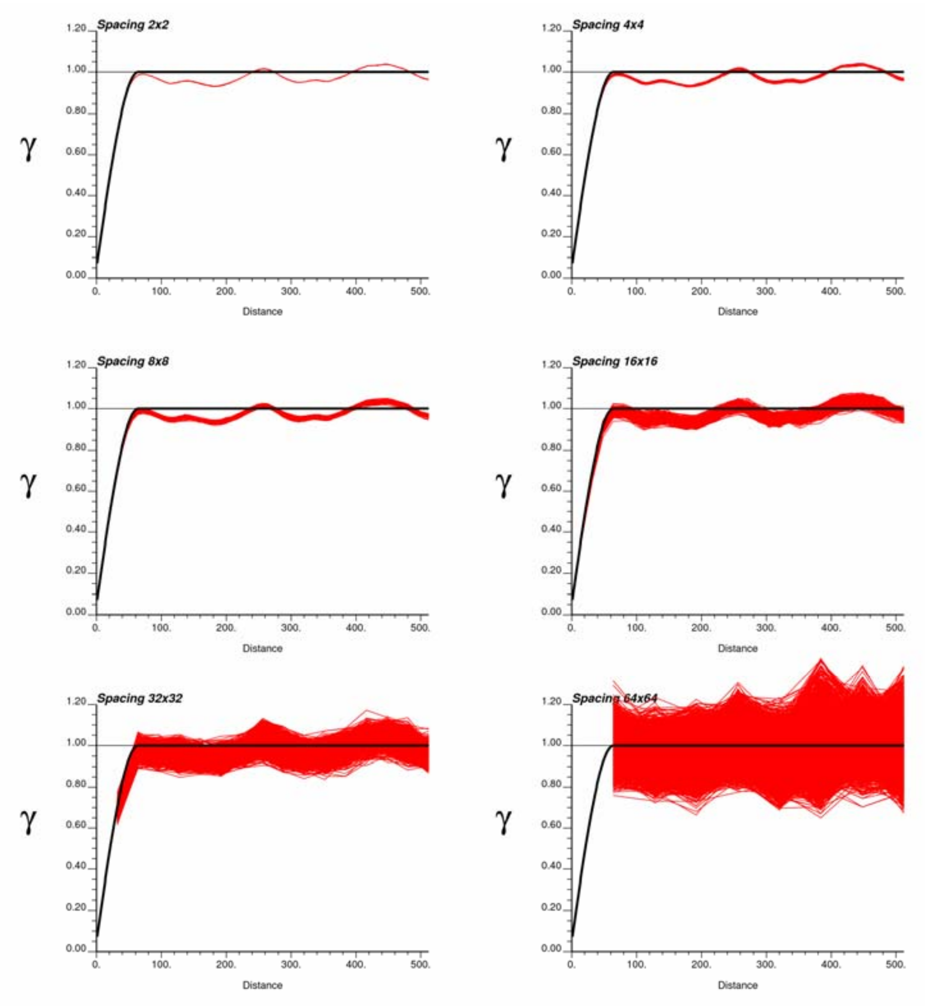

Figure 8: Calculated experimental variogram for 2x2, 4x4, 8x8, 16x16, 32x32 and 64x64. For each spacing there are m^2 different calculated variograms in each plot; the solid black line is the reference variogram model.

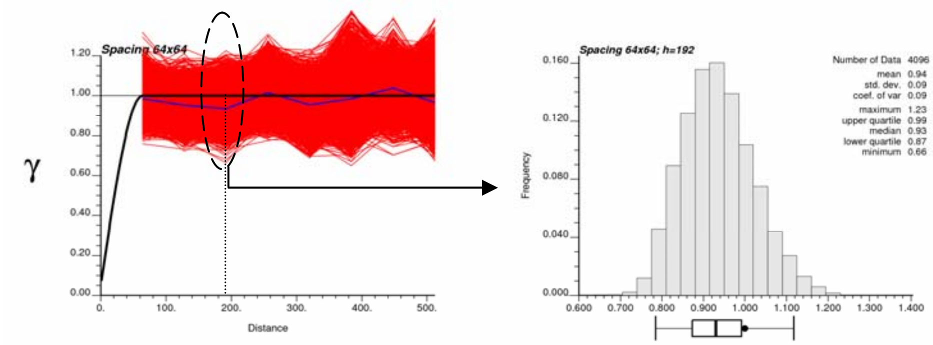

Figure 9: Distribution of the experimental variogram values at h=192 for 64x64 data spacing.

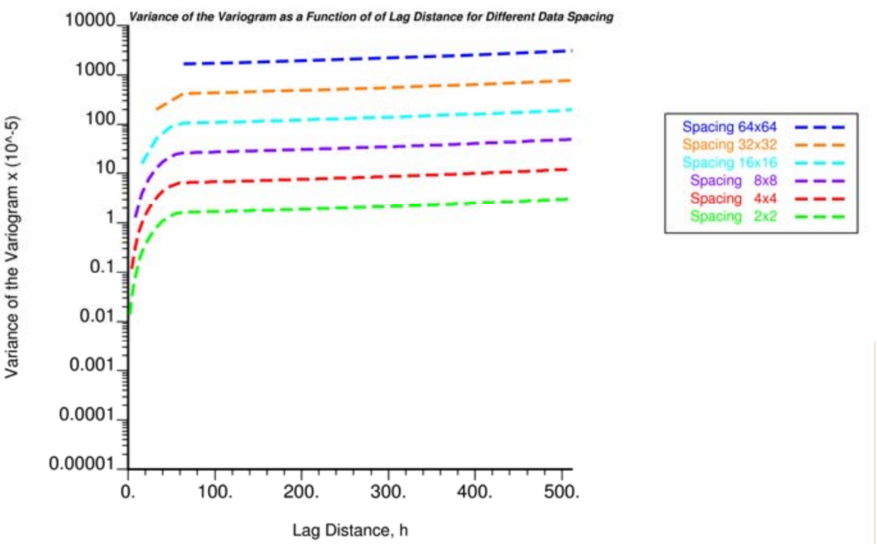

Figure 10: Variance of the variogram as a function of lag distance for different data spacing for Cressie (1985) Model.

Figure 11 shows the plot of the variance versus the lag distance, **h**, for the relevant data spacings in six different scenarios which are based on considering the variance of the variogram for: (1) the experimental points, (2) Cressie's model, and four possible modeling scenarios under a semi-automatic variogram fitting algorithm (VARFIT). The four scenarios involve modeling options related to the sill and the nugget effect to be fixed or variant: (a) fixed sill and fixed nugget, (b) fixed sill and variant nugget, (c) variant sill and fixed nugget, and (d) variant sill and variant nugget.

From Figure 11, we can see that the variance for the experimental variogram and the variance from Cressie's model have the same behavior for high lag distances. This behavior can also be seen in the last two figures for the case of variant sill-fixed nugget and variant sill-variant nugget. For the two cases of fixed sill (regardless of the nugget option), the variance of the variogram for the high lag distances is zero (and beyond a certain lag distance, cannot be shown in this semilog plot). These results are not surprising given that the impact of a variant sill will certainly yield a non-zero variance in the variogram beyond the range, while fixing the sill results in a distribution of the variogram (for **h**>a) that is a spike.

To gain a better appreciation for the variogram distributions and the impact of modeling options, Figure 12 shows the variogram distributions for these four modeling cases along with the experimental case is shown for a lag distance of 16 for the 8x8 data spacing. In general, all cases show an approximately normal distribution. The box plot reveals the impact of each case relative to the reference variogram model value at **h**=16 which is equal to 0.3988. We see that in all but the case of variant sill-variant nugget, the distribution of variogram values is systematically lower than the model value. Interestingly, this includes the experimental points.

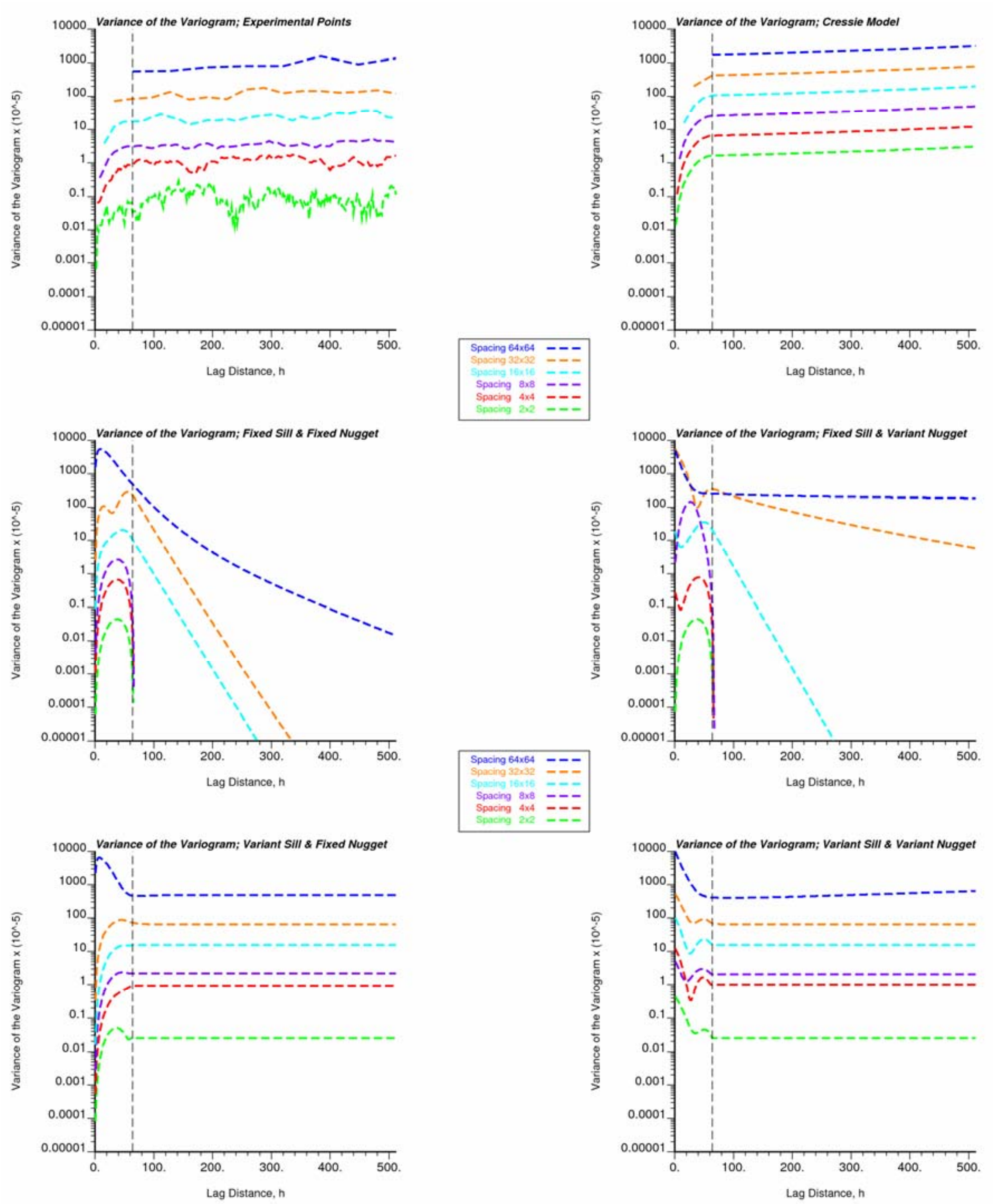

Figure 11: Variance of the variogram as a function of lag distance for different data spacing calculated for experimental variogram points (top left), for Cressie's (1985) model by using the reference model (top right) and for four different variogram models generated by VARFIT: fixed sill and fixed nugget effect (middle left), fixed sill and variant nugget effect (middle right), variant sill and fixed nugget effect (lower left), variant sill and variant nugget effect (lower right). The vertical dash line shows the range of the reference variogram.

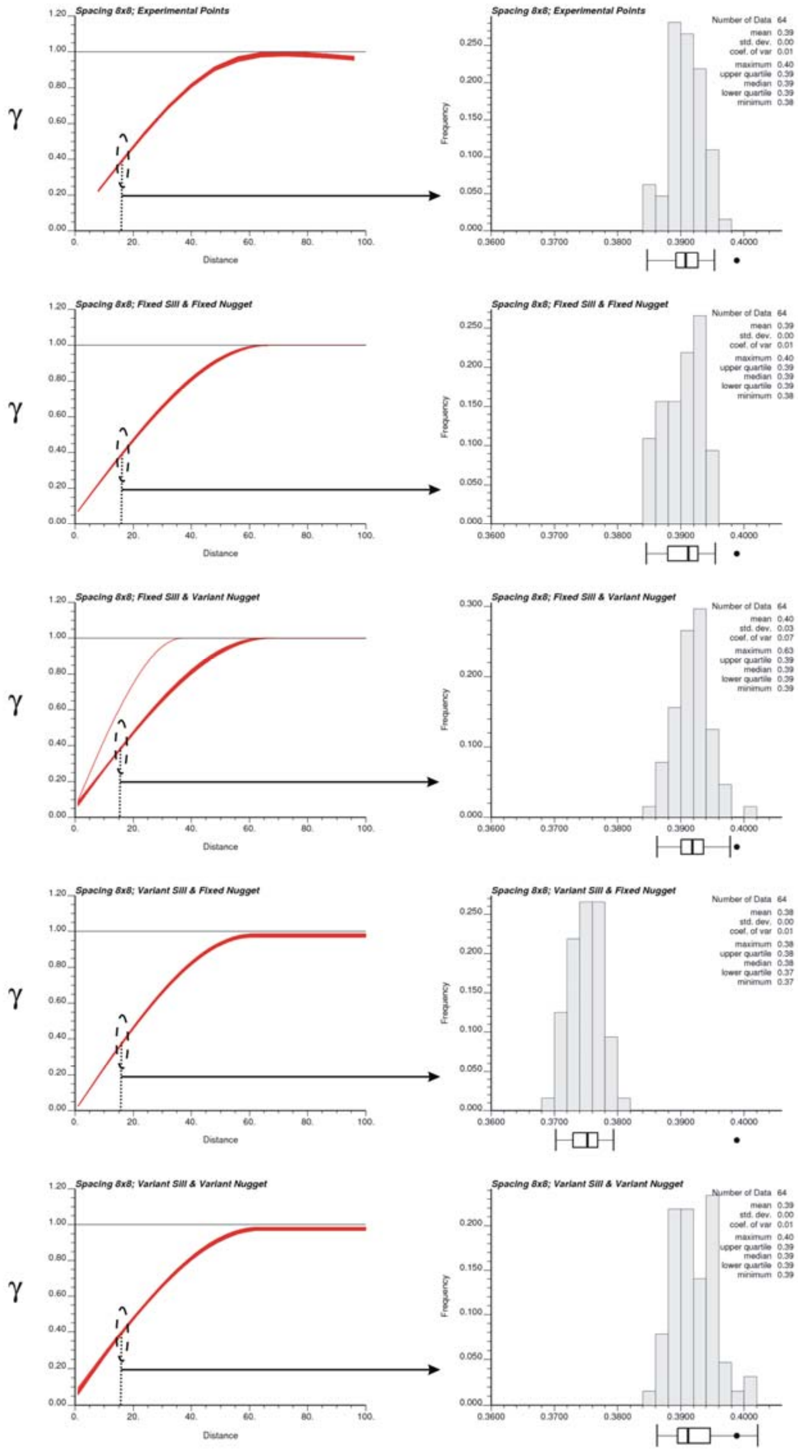

Figure 12: Histograms of variogram values at h=16 for 8x8 data spacing: experimental variogram (top row), fixed sill and fixed nugget (second row), fixed sill and variant nugget (third row), variant sill and fixed nugget (fourth row), variant sill and variant nugget (last row).

Figure 13 shows the plot of the variance of the variogram as a function of data spacing at fixed **h**=64 (the reference variogram range) in log-log plot for six different scenarios, the experimental variogram**,** the Cressie's model (the straight line), and four other cases that are generated by VARFIT. It can be seen from the plot that the variance of the variogram at each spacing from Cressie's model (the straight line) is systematically higher than the other five cases. Among these five cases, the experimental variogram and the two variant sill cases (regardless of nugget option) are virtually coincident with each other and are most similar to Cressie's model compared to the fixed sill cases.

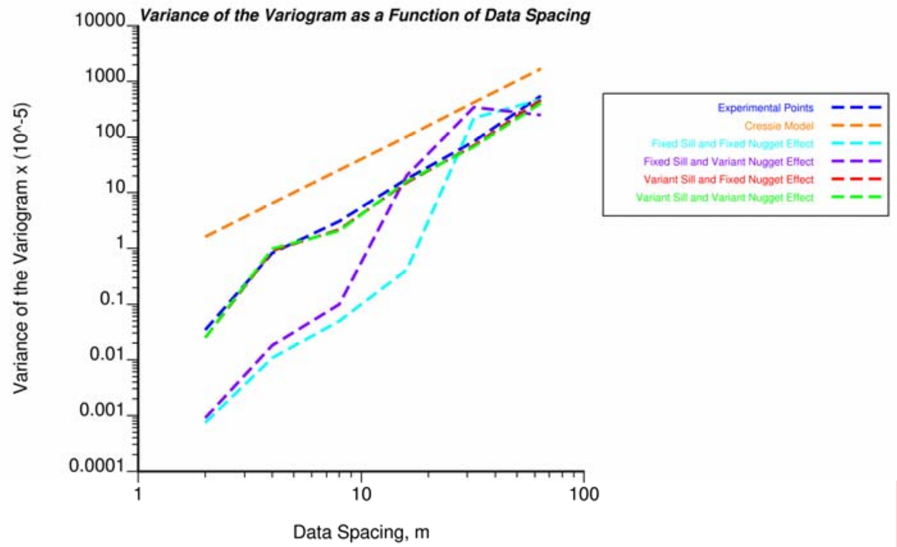

Figure 13: Variance of the variogram as a function of data spacing, *m*, at h=64 for six different scenarios: experimental variogram, Cressie's Model (the straight line), VARFIT with fixed sill and fixed nugget, VARFIT with fixed sill and variant nugget, VARFIT with variant sill and fixed nugget, VARFIT with variant sill and variant nugget).

Figure 14 shows the output of VARFIT for the four possible combinations of fixed sill or nugget and variant sill or nugget for the case of 32x32 data spacing. By fixing either nugget or sill, the variance of the variogram at **h**=0 and also at large values is equal to zero. In the case of fixing both sill and nugget, it is obvious that the uncertainty at lag distance of zero and at large lag distances are zero. For this case we can see that the variance of the variogram is artificially low for small and large lag distances **h**, and appears artificially high in between. For the case of fixed sill and variant nugget (the top right plot in Figure 14), allowing the nugget to vary permits great flexibility for the nugget to take on a large range of values in order to minimize the squared error in the overall fit, and as a result we see that this results in an artificially high variance at **h**=0. This is also evident in the fourth case where the sill is permitted to change along with the nugget. For the third case, fixed nugget and variant sill, again there is no variance at zero lag distance but as **h** increases the variance of the variogram is quite stable. For the last case where both the sill and nugget vary, the variance of the variogram is non-zero at both small and large lags. This last case is likely the most realistic as neither the nugget nor the sill can be predetermined. Despite this, we see that Cressie's model best approximates the case of fixed nugget and variant sill (see Figure 11).

To see the impact of the sill (first column of plots in Figure 14), we can examine the variance of the variogram as a function of the lag distance for the case of 16x16 and 32x32 data spacing (see

Figure 15). In both figures, the variance increases as **h** increases for small lag distances but when the sill is fixed the variance decreases dramatically near the reference range; if the sill is permitted to change, the variance function appears to flatten as it approaches the range.

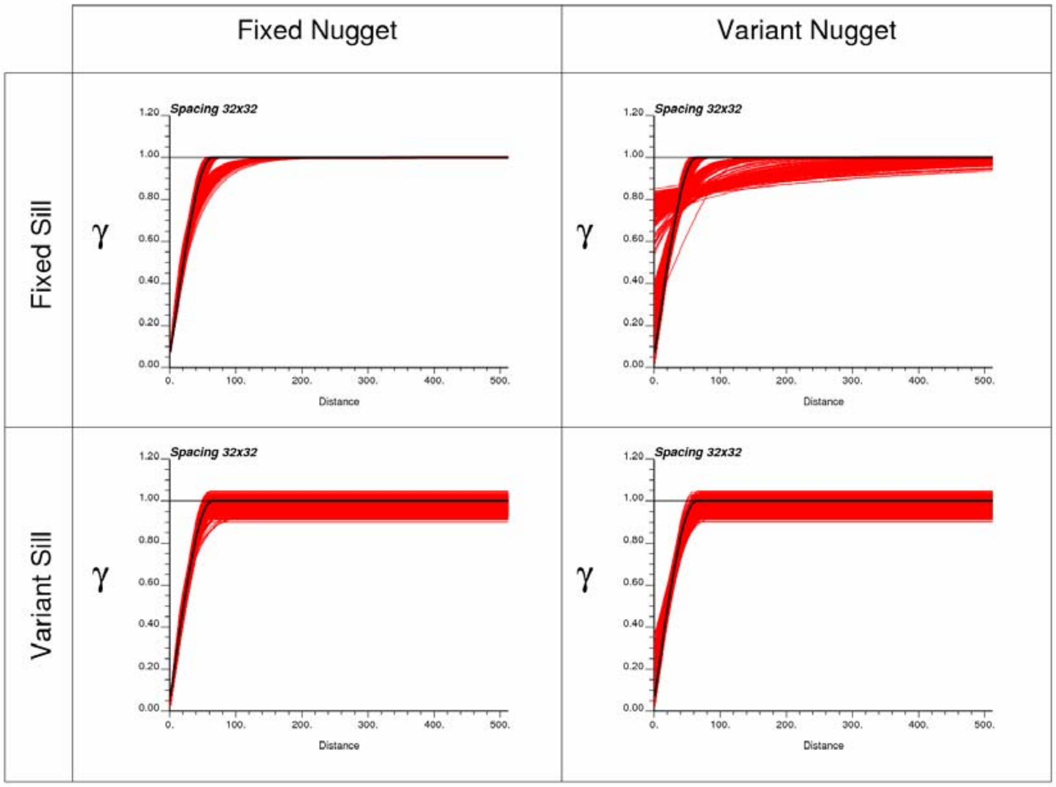

Figure 14: Variogram models for 32x32 by using VARFIT; There are four different cases: fixed sill and fixed nugget, fixed sill and variant nugget, variant sill and fixed nugget, variant sill and variant nugget; the solid thick line in black is the reference variogram model.

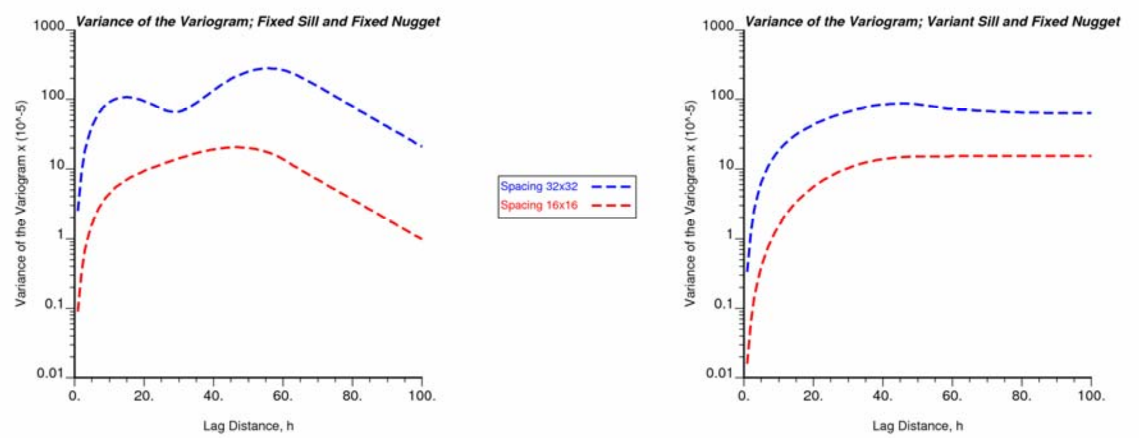

Figure 15: Plot of variance of the variogram versus lag distance, h, for 16x16 and 32x32 spacings; nugget effect in VARFIT is fixed in both plots, fixed sill (left) and variant sill (right).

Conclusions and Future Work

There are many issues in establishing a variogram model fit for geostatistical applications including clustered and noisy data and subjective choice of calculation and fitting parameters. This paper considered regularly gridded data and the impact of choices related to the calculation and fitting of experimental values. As the spacing of the data increases, the variance of the variogram also increases. Depending on the modeling choice, the variance in small lag distances can be much higher compared to large lag distances. A comparison with Cressie's approximation shows that Cressie's model is fairly accurate for the case of fixing the nugget effect and allowing the sill to vary; this is counter-intuitive to the real data scenario where neither parameters are usually known well enough to fix them beforehand. Nevertheless, for a fixed lag distance, the uncertainty in the variogram as a function of data spacing shows a similar increasing trend as Cressie's model.

This paper examined many different facets of variogram uncertainty, yet the areas for further exploration remain a relatively open field. This includes possible work in the following: (1) application of other theoretically derived models of variogram uncertainty, including a comparison to Ortiz and Deutsch's (2001) model; (2) consideration of clustered samples which would be most realistic, but this may be highly intractable; and (3) consideration of non-Gaussian random fields.

References

Bogaert, P. and Russo, D., "Optimal Spatial Sampling Design for the Estimation of the Variogram based on a Least Squares Approach", *Water Resources Research*, 35 (4), 1999, p. 1275-1289.

Chilès, J.P. and Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons Inc., New York, 1999, 695 pp.

Cressie, N., "Fitting Variogram Models by Weighted Least Squares", *Mathematical Geology*, 17 (5), 1985, p. 563-586.

Cressie, N., *Statistics for Spatial Data*, John Wiley & Sons Inc., New York, 1991, 900pp.

David, M., *Geostatistical Ore Reserve Estimation*, Elsevier, Amsterdam, 1977.

Davis, B.M. and Borgman, L.E., "Some Exact Sampling Distributions for Variogram Estimators", *Mathematical Geology*, 11 (6), 1979, p. 643-653.

Davis, B.M. and Borgman, L.E., "A Note on the Asymptotic Distribution of the Sample Variogram", *Mathematical Geology*, 14 (2), 1982, p. 189-193.

Deutsch, C.V. and Journel, A.G., 1998: *GSLIB - Geostatistical software library and users guide*. Oxford University press, 2nd Edition.

Genton, M.G., "Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure", *Mathematical Geology*, 30 (4), 1998, p. 323-345.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, 1997, 483pp.

Gringarten, E. and Deutsch, C.V., "Teacher's Aide: Variogram Interpretation and Modeling", *Mathematical Geology*, 33 (4), 2001, p. 507-534.

Journel, A.G., *Fundamentals of Geostatistics in Five Lessons*, Volume 8 Short Course in Geology, American Geophysical Union, Washington, DC, 1989, 40 pp.

Larrondo, P. F. and Neufeld, C. T. and Deutsch, C.V., "VARFIT: A Program for Semi-Automatic Variogram Modeling", *Centre for Computational Geostatistics Annual Report 5*, Department of Civil and Environmental Engineering, University of Alberta, 2003, 17 pp.

Matheron, G. *La Théorie des Variables Régionalisée et ses Applications*, Masson, Paris, 1965.

Neufeld, C. T. and Deutsch, C. V., "Developments in Semiautomatic Variogram Fitting", *Centre for Computational Geostatistics Annual Report 6*, Department of Civil and Environmental Engineering, University of Alberta, 2004, 12 pp.

Omre, H., "The Variogram and its Estimation", in Verly, G., David, M., Journel, A.G. and Marechal, A., eds., *Geostatistics for Natural Resources Characterization*, D. Reidel Publishing Company, Dordrecht, 1984, pp. 107-125.

Ortiz, J.M. and Deutsch, C.V., "Calculation of Uncertainty in the Variogram", *Mathematical Geology*, 34 (2), 2001, p. 169-183.

Pardo-Igúzquiza, E. and Dowd, P., "Variance-Covariance Matrix of the Experimental Variogram: Assessing Variogram Uncertainty", *Mathematical Geology*, 33 (4), 2001, p. 397-419.

Switzer, P., "Inference for Spatial Autocorrelation Functions", in Verly, G., David, M., Journel, A.G. and Marechal, A., eds., *Geostatistics for Natural Resources Characterization*, D. Reidel Publishing Company, Dordrecht, 1984, pp. 127-140.