Selected Implementation Issues with Sequential Gaussian Simulation

Stefan Zanon (szanon@ualberta.ca) and Oy Leuangthong (oy@ualberta.ca)
Department of Civil & Environmental Engineering
University of Alberta

Abstract

Sequential Gaussian simulation (SGS) is a common method used to create multiple equiprobable numerical models based on some conditioning data. The basic methodology is straightforward, but some implementation details are important for successful results. This paper explores some of the issues associated with the implementation of SGS and the reasoning behind choices such as: simulation path, search strategies, number of conditioning data, some effects of secondary data, and ergodic fluctuations.

Introduction

Since the early 1990's, SGS [5] has gained popularity in practice due to its simplicity, flexibility, and reasonable CPU time [1]. SGS is a simulation algorithm based on kriging where by more variability is injected into the model. The SGS work-flow can be described in five basic steps:

1. Choose the stationary domain and transform the data to a Gaussian distribution.
2. Define a path to visit every location.
3. At each location:
   (a) search to find nearby data and previously simulated values,
   (b) calculate the conditional distribution, and
   (c) perform Monte Carlo simulation to obtain a single value from the distribution.
4. Repeat step 3 until every location has been visited.
5. Transform the data and all simulated values back to their original distribution.

This will produce one possible realization. More realizations can be created by changing the random number seed.

The theory behind SGS has been explained numerous times [1, 2, 4], but it is the details behind these steps that warrant further explanation. In programs like SGSIM [2], many assumptions are made about the data that must be satisfied or the final results may be erroneous. Understanding these assumptions and their affect on the resulting realizations will help users find deficiencies in their models and/or identify other potential problems.

Data and Transformation

Before simulation can be performed the model area must be defined and the input data identified. In general, the data must come from a single underlying distribution. For geological data, this amounts to considering only data from the same rock type with similar properties. Under the assumption of stationarity [6], the mean, variance, and higher order statistics are assumed constant throughout the area, that is, $E\{\mathbf{u}\} = m$ and $\sigma^2(\mathbf{u}) = \sigma^2$.

Sometimes the data will violate the assumption of stationarity and the mean or variance will change with location. Typically these changes are smooth and there are no clear boundaries between the high and low value regions. In these situations, trends in the attribute should be modelled [1]. This trend can be used in one of two ways: remove the trend from the data and work with the stationary residuals or treat the trend as secondary data and use a specialized form of kriging to account for this information.

SGS is designed to work with input data that follows a Gaussian distribution. The convention is to use the standard normal distribution, N(0,1). Rarely will the input data fit this distribution exactly. Data transformation is then required to convert the input distribution to the standard normal distribution. Before transformation can be performed, the cumulative distribution function (cdf) for the input distribution and target distribution must be known. These two cdfs are related by a one-to-one quantile transform; thus, the transform is reversible and the simulated values can be returned to their original units. A problem can arise when many data have the same constant value. This results in a spike or vertical jump in one of the cdfs. This causes a problem in the transformation. A quantile in one cdf is equal to a range of quantiles in the other, so the transform is no longer one-to-one. To correct for these jumps, despiking [1] is required to break the ties in the data.

For simulation to be accurate, a good understanding of the original distribution is required. If no other information is available, the cdf will be created directly from the input data. This data is not typically representative of the global distribution since this is not the goal of the sampling process. If a reference distribution is available, it can be used as the target distribution and the data is transformed relative to this distribution. If this is not possible, declustering [2] can be used to try and correct for non-representative sampling. Declustering techniques use different methods to account for the closeness of the surrounding data and then weights the data based on their spatial spread. The histogram is then altered by accounting for the weights when calculating the distribution.

When the study considers multiple correlated variables for SGS, each variable must be Gaussian. Typically, each variable is independently transformed to a normal distribution, but this only ensures univariate Gaussianity. For correlated variables, the multivariate relationships must be accounted for in the transformation process, otherwise the relationships will not be preserved. To achieve this, an alternative transformation process like stepwise conditional transformation [9, 10] should be considered. Stepwise transformation creates independent, multi-Gaussian variables that can be independently simulated. The relations between the variables are preserved in the back-transformation process.

Simulation Path

Sequential Gaussian simulation makes no assumptions as to the order in which the unsampled locations are visited. However, previously simulated values will be used as conditioning data , therefore, the order can influence the model. To minimize this influence, a random starting location and path have been found to produce the least effect on the model over multiple realizations. Alternatives, such as the regular path and spiral path [11], have been considered but any perceived benefits in CPU efficiency or input data propagation comes at the cost of variogram reproduction.

The random path can suffer from the preferential use of nearby data, resulting in poor reproduction of the long range variogram features. To avoid this problem, a multiple grid search [13] can be incorporated. A coarse grid, relative to the final grid size, is first defined and simulated randomly. This grid size is then reduced in several increments until the final model grid has been reached. As the grid size reduces so will the distance to the conditioning data. In this way the long range features are first captured through the coarse grid and each refinement in the grid will reproduce features at a shorter scale. The multiple grid search used in conjunction with the random path will improve variogram reproduction with almost no increase in CPU time [13].

Searching for Local data

Before kriging can be implemented, the surrounding data must be identified. The search is limited by a search radius in each principle direction. These distances should equal or exceed the variogram range to ensure adequate variogram reproduction. The data beyond the range will provide limited information to the kriging estimate.

It is common practice to assign the input data to the grid nodes. This has the advantage that input data at the nodes will be reproduced exactly in the final model. The disadvantage is that only the closest data is retained inside of each grid cell. Data cannot cross the cell boundaries so the rest of the data inside the cell and any data beyond the modelling area is lost; however, they will be used to define the input statistics. The input data and previously simulated values now fall on the same regular grid system; this allows the covariance between any two correlated locations to be conveniently stored in a look-up table. The spiral search [2] utilizes this look-up table to develop a search path through the surrounding data based on the covariance to the point being simulated. Starting at the simulated node, the most correlated nodes are checked first. The search continues to spiral through increasingly less correlated locations until the search radius is reached or the required number of data are located.

If the input data are not assigned to the grid system then only previously simulated values can be located using the spiral search. A second search, called super block search [2], is required to locate nearby input data. This search starts by superimposing a coarse grid over the modelling area to create blocks. These super blocks are usually much larger than the simulation grid and are independent of the model. Starting at one corner, the blocks are numbered and checked for input data. An array is built that keeps track of the cumulative number of data that have been located in all preceding blocks. The data is placed in a second array in the order they are identified. These two arrays can then be used

to identify the number of data in any block and their corresponding location in the data array. A template is constructed based on the number of super blocks that encompass the search radius. When an unsampled location has been chosen, the super block containing this location is identified and the template is centred on the block. This will identify all super blocks that must be checked for conditioning data and this data is quickly assembled. An exhaustive search is then used to calculate the covariances and the closest data are identify.

The above search routines are only concerned with identifying the most correlated data to the point being simulated. However, a good kriging estimate should consider data from every direction. A classical example of this problem is drillhole data, where an estimate can easily be based on data from only one drillhole. The vertical sample spacing of drillhole data is usually several orders of magnitude smaller than the horizontal spacing. These data are highly correlated to each other and the resulting estimate will be biased to one drillhole. An octant search can be used to avoid this problem. The octant search divides the surrounding 3D area into equal octants. A search is performed and only a maximum number of data is considered from each octant. This forces the data to come from different directions at the expense of ignoring closer, but more redundant, data.

Kriging

The theory behind SGS is based on using every previously simulated value and input data throughout the simulation process [5]. In practice, only the closest conditioning data are used, up to a maximum number, to keep CPU time reasonable. The reasoning behind this decision is that the closest data will screen data that are further away and the additional information is deemed small enough to ignore [5]. This assumes the data comes from the same stationary population. The choice of the maximum number is linked to two things: the speed required to generate a realization, and the accuracy of the kriged estimate and variance.

The kriging step in simulation consists of locating n conditioning data, inverting an n x n covariance matric, and multiplying the inverted matrix by another n x 1 covariance matrix. As n increases, the CPU requirements are proportional to n, n^3, and n^2, respectively. For small values of n the CPU requirements will be a combination of all three steps, but as n increases in size, the CPU time will be dominated by the inversion step. For example, a 100 x 100 grid was simulated using 300 randomly placed input data and the CPU time was tracked as the maximum number of conditioning data were varied between 5 and 300 (Figure 1). Initially, the change in CPU time is small since other operations still dominate the time required to complete a full simulation. As the number of data increases, the CPU requirements approach a slope of 3 in log-log scale, as represented by the solid line, and CPU demands increase rapidly. This example was designed to use the specified maximum number of conditioning data at every location.

The kriging system also provides a measure of uncertainty in the estimate by way of the kriging variance. This variance is minimized for any set of conditioning data, but this variance is reduced through the addition of more data. The change in the kriging variance by adding a single datum can be predicted, within a bounding range, when the least informative datum is removed [3].

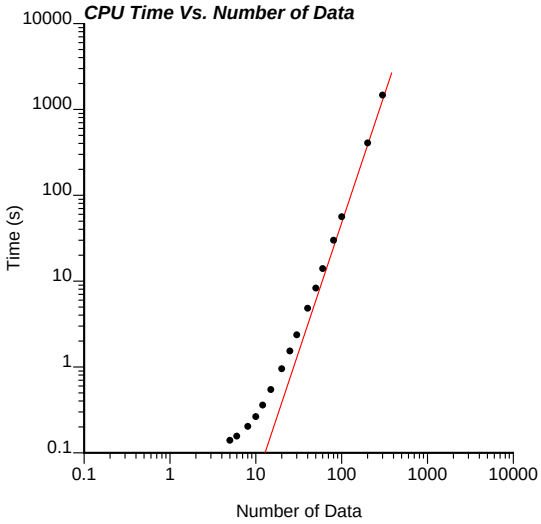

Figure 1: The CPU time required to simulate a 100 x 100 grid, starting with 300 randomly placed input data, and the number of conditioning data are varied between 5 and 300. The line is representative of n^3 or a slope of three in log-log scale.

Starting with n data the kriging weights, λ_α, and variance, $\sigma^2_{sk,n}$, are calculated. The location with the smallest absolute kriging weight, k, is then found and removed. A second kriging is performed using the remaining $n-1$ conditioning data and a new set of kriging weights are found, λ'_α:

$$\sum_{\alpha=1,\alpha\neq k}^{n} \lambda'_\alpha \cdot C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}, \mathbf{u}_\beta),\ \beta = 1, \ldots, n,\ \beta \neq k \tag{1}$$

where the number of conditioning data is $n-1$ but the location of k is any one value from 1 to n. This second set of kriging weights, λ'_α, will produce a slightly higher kriging variance, $\sigma^2_{sk,n-1}$.

Moving to location k, a third set of kriging weights, $\hat{\lambda_\alpha}$, are calculated using the same set of $n-1$ data as above:

$$\sum_{\alpha=1,\alpha\neq k}^{n} \hat{\lambda_\alpha} \cdot C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}_k, \mathbf{u}_\beta),\ \beta = 1, \ldots, n,\ \beta \neq k \tag{2}$$

It is now possible to predict the initial kriging weights, λ_α, by using the λ'_α and $\hat{\lambda_\alpha}$ kriging weights:

$$\lambda_\alpha = \lambda'_\alpha - \hat{\lambda_\alpha} \cdot P_n,\ \alpha = 1, \ldots, n, \alpha \neq k \tag{3}$$

where the k weight can not be predicted and the variable P_n is based on the following relationship:

$$P_n = \frac{P'_n}{C(\mathbf{u}_k, \mathbf{u}_k) - \sum_{\alpha=1,\alpha\neq k}^{n} \hat{\lambda_\alpha} \cdot C(\mathbf{u}_k, \mathbf{u}_\alpha)} \tag{4}$$

where $C(\mathbf{u}_k, \mathbf{u}_k) = 1$ when the data is standard normal, weights are from the third kriging and P'_n is equal to:

$$P'_n = C(\mathbf{u}, \mathbf{u}_k) - \sum_{\alpha=1, \alpha \neq k}^{n} \lambda'_\alpha \cdot C(\mathbf{u}_k, \mathbf{u}_\alpha) \tag{5}$$

where the weights come from the second kriging.

The difference between the $\sigma^2_{sk,n-1}(\mathbf{u})$ and the $\sigma^2_{sk,n}(\mathbf{u})$ variances can be predicted within a bounding region based on P_n and P'_n. The following inequality provides the limits for the upper and lower bounds [3]:

$$(P_n)^2 \leq \sigma^2_{sk,n-1}(\mathbf{u}) - \sigma^2_{sk,n}(\mathbf{u}) \geq (P'_n)^2 \tag{6}$$

The above prediction is performed on a small example in shown in Figure 2. Starting with $n = 5$ the kriging variance is calculated as $\sigma^2_{sk,n}(\mathbf{u}) = 0.3097$. Removing the fifth datum, the least informative, a second kriging is performed yielding a slightly higher kriging variance at $\sigma^2_{sk,n-1}(\mathbf{u}) = 0.3112$. Shifting the location of interest to the fifth data, a third kriging is performed. The last two sets of kriging weights are then applied to Equations 4 and 5 to calculate (P_n) at -0.0799 and (P'_n) at -0.0195. The upper and lower bounds for the variance change are then calculated to be 0.0064 and 0.0004, respectively. The true change in the variance is 0.0015 and this falls with in the bounding limits.

In application, the calculation of three sets of kriging weights at every location is impractical. To provide some better guidelines, three locations were chosen to track how the accuracy of the kriging estimate and variance changed as the number of conditioning data increased. Starting with 100 randomly placed data, three locations were chosen (Figure 3). At each location the kriging estimate and variance were tracked as the number of conditioning data was varied between 1 and 100 (Figure 4). The best possible values are shown as dotted lines and occur when all available conditioning data are used. Initially, the kriging estimate and variance, for each location, show a large degree of deviation for the best value. As the number of conditioning data increased to 8 or 10, both the estimate and variance begin to converge. Past 10 data, both the estimates and variances show only a limited degree of improvement.

Kriging in SGS provides an estimate and variance for every unsampled location. To create a simulated value a random value is drawn from a residual distribution with a mean of zero and a variance equal to the local kriging variance. This residual value is added to the kriged estimate to produce a simulated value. If few conditioning data are used, then the estimate will be poor and the variance high. This gives rise to the potential to draw simulated values outside the plausible range. As more conditioning data are used the estimate will be improved and the kriging variance is reduced. The reduced variance will limit the possibility of drawing implausible values.

Accounting for Secondary Data

The possibility of drawing outside the plausible range is also a problem when secondary data are incorporated into the process. Simple kriging can be applied with a locally varying mean (LVM). A secondary source of information is used to predict the mean over the modelling area. When the kriging estimate is close to the global mean of zero, in normal space, the

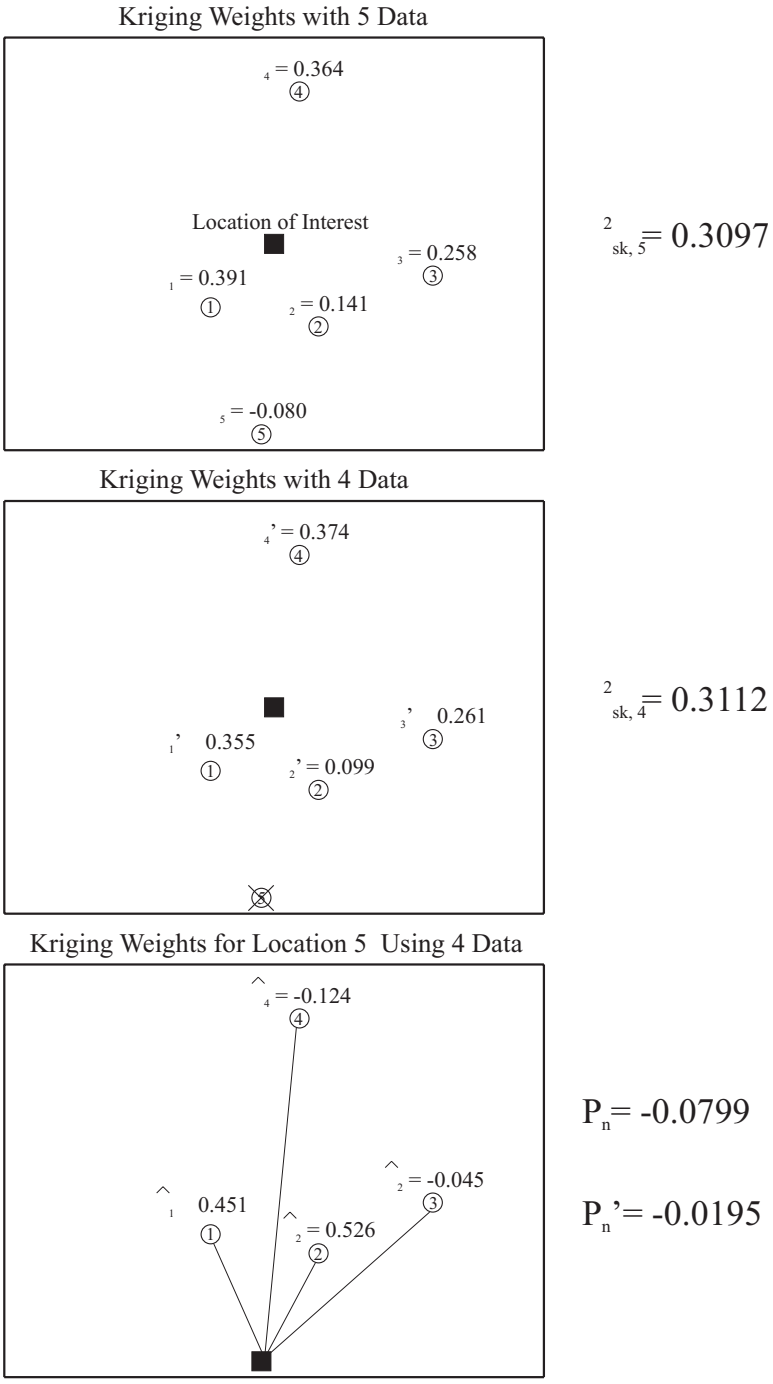

Figure 2: The top figure shows the location, kriging weights and kriging variance for the five conditioning data. The middle figure shows the kriging weights and variance for the second kriging when the fifth point is excluded. The bottom figure shifts the point of interest to the fifth data point and a third kriging is performed. Using the second and third sets of kriging weights the P_n and P'_n values are calculated.

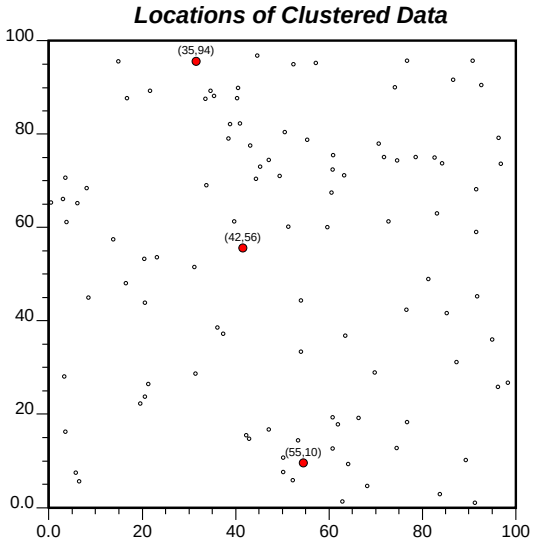

Figure 3: Location of the three points of interest (large dots) and the 100 input data in the 100 x 100 example.

kriging variance can take on any value between 0.0 and 1.0. As the estimate deviates away from the global mean, a large variance increases the probability of drawing a value beyond the expected range (Figure 5). These outliers can cause the global variance to increase and this causes problems in the back-transformation.

A variance inflation problem can also be seen with collocated cokriging (CCK) [14]. CCK is a simplified version of full cokriging that uses only the collocated secondary data and assumes an intrinsic model of coregionalization [8]. Removing the need to model the linear model of coregionalization greatly simplifies the simulation process, but limiting the number of secondary data causes the kriging variance to be higher than expected. This increase in variance at every unsampled location causes the global variance to be inflated and again problems arise in the back-transformation.

To fix the variance inflation problems in LVM and CLK a Self-Healing algorithm [12] was developed. Self-Healing is a dynamic correction factor that is applied to reduce the global variance. As simulation progresses, the algorithm tracks the global variance and identifies problem locations. At these locations, a correction factor between 0.0 and 1.0 is calculated based on the severity of the problem. This factor is then applied to the local variance and translates to an improved global variance. Self-healing does not affect the expected value of the local distribution and is only applied when the global variance is inflated.

Ergodic Fluctuations

Statistical fluctuations are inherent in simulation; however, the fluctuations should be reasonable and unbiased. Simulation in normal space produces simulated values that are

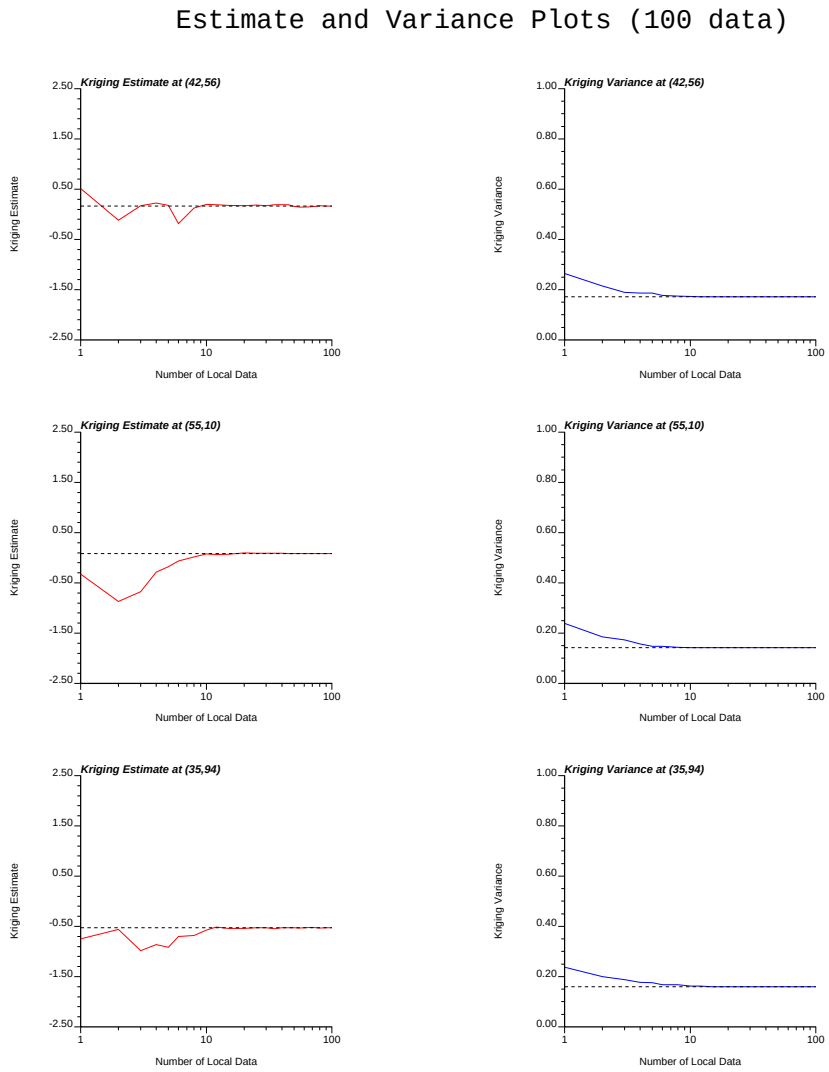

Figure 4: The top two plots shows the change in the kriging estimate (left) and variance (right) for location (42,56) as the number of conditioning data is varied. The middle plot shows the results for location (55,10) and the bottom plots are for location (35,94). The dotted line on each plot is the best possible estimate when 100 conditioning data are used.

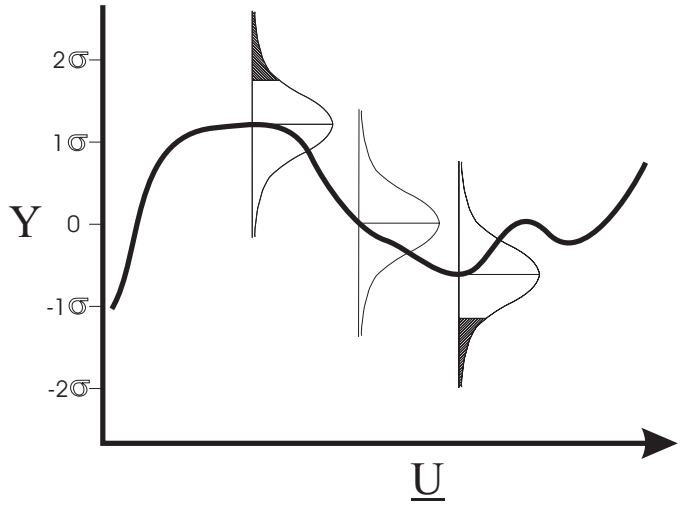

Figure 5: Illustration of a locally varying mean (solid line) versus location, **u**, and three conditional distributions. The shaded regions of the two outside distributions cause variance inflation because there is a too high probability to draw large and small values.

approximately standard normal in expected value. For any one realization, minor fluctuations from a zero mean and unit variance are expected; however, when these values are back-transformed to original units a slight shift of the mean in normal space may translate to a more significant shift of the mean in original units. This problem is compounded when shifts in the variance are incorporated. This is particularly true for skewed distributions, which is the case for most geological variables.

Deviations from the limit standard normal distribution could be due to a number of factors. Firstly, the algorithms employed are based on an assumption of stationarity. Simulation of non-stationary data can lead to shifts in the mean and/or variance of simulated values in normal space. Secondly, SGS assumes the data are multiGaussian in a spatial context. There are no techniques to ensure multiGaussianity in the spatial domain.

For example, consider the positively skewed distribution and its corresponding normal scores shown in Figure 6. The effect of deviations from the standard normal distribution can be assessed by generating *approximately* standard normal distributions and back-transforming the data to original units. For this exercise, a series of deviations of the standard normal distribution were tested: 17 different standard deviations ranging between 0.80 to 1.2 at intervals of 0.025. For each standard deviation, nine different mean values were assigned ranging from -0.1 to 0.1 at intervals of 0.025. Each Gaussian distribution is then back-transformed to the original units and the corresponding mean and variance in original space is calculated.

In total, this yields 153 conditioning data for the global mean and standard deviation in original space, the two variables of interest. This data was then kriged to produce maps of how the two variables are affected in original space when the variables deviate in normal space, Figure 7. The change in the mean in original space is most pronounced when the normal space mean and standard deviation is greater than the standard (0,1). Alternatively,

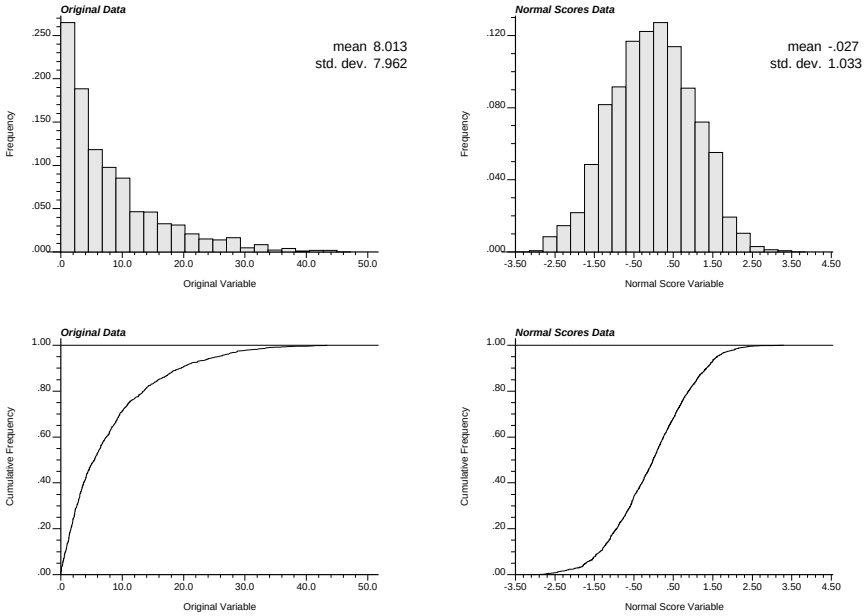

Figure 6: Distribution of original data (left) and its corresponding normal scores (right). Histograms (top) and cumulative histograms (bottom) are shown for both.

when the normal space mean and standard deviation is both lower than the standard (0,1), a region of low values appears in the lower left corner; this region may be due to edge effects in kriging.

The estimated map of standard deviation in original space shows similar features as the original space mean. The region of high values is larger than that observed in the estimated mean, while the edge effects are more apparent at low values of the normal space standard deviations. Overall, this small exercise shows the sensitivity of the original space summary statistics to the ergodic fluctuations inherent in stochastic simulation in normal space.

To mitigate the effects of fluctuations in normal space and its translation to original space of the data, a standard transform can be applied to the simulated values to ensure reproduction of the histogram and its corresponding summary statistics [7]. The transform is applied over individual realizations to ensure reproduction of the global histogram for each realization. Alternatively, "sets" of realizations can also be transformed and data would still be honoured; however, this does not guarantee that the global histogram per realization is reproduced.

Final Remarks

Once the modelling process has been completed, each realization should be checked for problems. The things to look at include, but are not limited to, the following list: basic statistical parameters, histogram, variogram, data reproduction and geological structure.

1. In original units, the global mean and variance should be close to the input values for

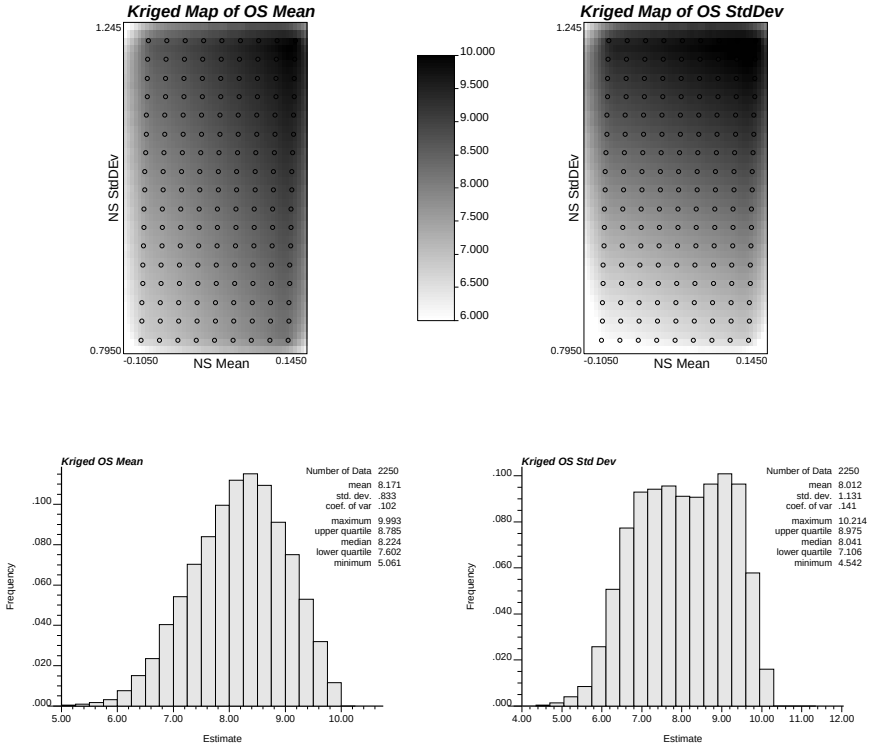

Figure 7: Effect of deviations of the mean and standard deviation in normal space on the mean (Top Left) and standard deviation (Top Right) in original space. The corresponding histograms of the mean and standard deviation are shown on the bottom.

each realization and for the average of all realizations.

2. The global histogram should follow the same shape as the input histogram and any extreme values should be checked.

3. The output variogram should reproduce the model variogram. This includes the nugget, short range and the long range structure, and the sill.

4. When the data are assigned to the grid, they should be reproduced in every realization.

5. The model should be visually inspected to check that the basic geological structure is honoured.

Ensuring that each of these checks are performed and satisfied will help to catch major errors in the resulting model. If the model fails a check some time must be spent to either identify the problem or to determine if it is acceptable under the given conditions and data.

References

[1] C. V. Deutsch. *Geostatistical Reservoir Modeling*. Oxford University Press, New York, 2002.

[2] C. V. Deutsch and A. G. Journel. *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, New York, second edition, 1998.

[3] L. S. Gandin. Objective analysis of meteorological fields, 1963. Translated from Russian: Israel Program for Scientific Translations, Jerusalem, Israel, 1965.

[4] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.

[5] E. H. Isaaks. *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*. PhD thesis, Stanford University, Stanford, CA, USA, October 1990.

[6] A. G. Journel and Ch. J. Huijbregts. *Mining Geostatistics*. Academic Press, New York, 1978.

[7] A. G. Journel and W. Xu. Posterior identification of histograms conditional to local data. *Mathematical Geology*, 26:323–359, 1994.

[8] O. Leuangthong. Short note on models of coregionalization. In *Centre for Computational Geostatistics Report Four*, Edmonton, Alberta, March 2002. University of Alberta.

[9] O. Leuangthong. *Stepwise Conditional Transformation*. PhD thesis, University of Alberta, Edmonton, AB, Canada, June 2003.

[10] O. Leuangthong and C. V. Deutsch. Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*, 35, 2003.

[11] J. A. McLennan. The effect of the simulation path in sequential gaussian simulation. In *Centre for Computational Geostatistics Report Four*, Edmonton, Alberta, March 2002. University of Alberta.

[12] T. Faechner S. Zanon and C. V. Deutsch. Improved integration of secondary data using self-healing sequential gaussian simulation. In *Application of Computers and Operations Research in the Mineral Industry: Proceedings of the 30th International Symposium*, Phoenix, AZ, USA, 2002. Society for Mining, Metallurgy, and Exploration.

[13] T. T. Tran. Improving variogram reproduction on dense simulation grids. *Computers and Geosciences*, 20(7/8):pg. 1161–1168, 1994.

[14] W. Xu, T. T. Tran, R. M. Srivastava, and A. G. Journel. Integrating seismic data in reservoir modeling: The collocated cokriging alternative. *Society of Petroleum Engineers*, 1992. SPE 24742.