The Transfer of Uncertainty in the Mean through Simulation

Martha E. Villalba and Clayton V. Deutsch

A methodology is developed to transfer the uncertainty in the mean through simulation to obtain a more complete evaluation of uncertainty. The paper demonstrates the use of a simple methodology to account for parameter uncertainty. Multiple distributions are constructed and used in simulation as reference distributions. Original values are transformed into Gaussian units according to a specified reference distribution. The uncertainty in the mean of the univariate distribution is accounted for by changing the reference distribution for transformation.

1. Introduction

Geostatistical simulation is usually performed with a fixed input distribution; this fixed univariate distribution comes from the input data and assumes a mean without uncertainty. An important aspect of this thesis is that uncertainty in the mean of the input distribution (UMID) must be transferred through simulation for a more complete understanding of uncertainty. The techniques of conditional finite domain (CFD) and stochastic trend (ST) provide the UMID. Multiple distributions could be constructed and used in geostatistical simulation as reference distributions.

Simulation is performed in Gaussian space because a consistent multivariate distribution is required and the multivariate Gaussian is the only known practical multivariate distribution. Original values are transformed into Gaussian units according to a specified reference distribution. The uncertainty in the mean of the univariate distribution is accounted by changing the reference distribution for transformation. A sequential Gaussian simulation (SGS) algorithm is adopted in this thesis; however, any Gaussian algorithm for simulation could be used. SGS is used because it is simple, flexible, and reasonably efficient (Deutsch, 2002).

A change in local and global uncertainty is expected when UMID is transferred through the simulation process. A measure of local uncertainty is available at every location by generating a set of *L* realizations:

$$z^{(l)}(\mathbf{u}), l = 1,\ldots,L$$

$$F(\mathbf{u}; z \mid (n)) = \text{Prob}\{Z(\mathbf{u}) \le z \mid (n)\} \tag{1}$$

Local uncertainty could be used for planning and decision making; however, some applications require the uncertainty of more than one location simultaneously, a measure of the joint uncertainty about attribute values at several locations taken together. This spatial uncertainty is modeled by generating multiple realizations of the joint distribution of the attribute value (Goovaerts, 1997). Those realizations should reasonably reproduce the sample histogram and the semivariogram model. The set of simulated maps is generated by sampling the *N*-variate ccdf that models the joint uncertainty at the *N* locations $\mathbf{u}'_j$:

$$\left\{z^{(l)}(\mathbf{u}'_j), j = 1,\ldots,N\right\}, l = 1,\ldots,L$$

$$F(\mathbf{u}'_1,\ldots,\mathbf{u}'_N; z_1,\ldots,z_N \mid (n)) = \text{Prob}\{Z(\mathbf{u}'_1) \le z_1,\ldots,Z(\mathbf{u}'_N) \le z_N \mid (n)\} \tag{2}$$

Spatial uncertainty is a result of our incomplete knowledge of the spatial distribution of the variable of interest.

2. Methodology

The probability distributions of continuous data are often summarized by a central value such as the mean (Deutsch, 2000). The mean of the distribution is a fixed parameter in the simulation process. Where there are *n* data values $z(\mathbf{u}_i)$, $i = 1,\ldots,n$ with different weights $w(\mathbf{u}_i)$, $i = 1,\ldots,n$:

$$m_r = \frac{\sum_{i=1}^{n} z(\mathbf{u}_i) w(\mathbf{u}_i)}{\sum_{i=1}^{n} w(\mathbf{u}_i)}$$

The uncertainty in the mean of the original distribution (σ_m) comes from one of the techniques such as the CFD or ST. The calculated fluctuations of the mean are summarized by distributions that have different mean values. The number of distributions or reference distributions in the simulation could be defined by L equally spaced quantiles:

$$p_l = \frac{l-0.5}{L}, \quad l = 1,...,L$$

The specific mean values corresponding to these quantiles are computed from a non standard Gaussian distribution computed to represent the UMID like standard deviation.

$$y_l = G^{-1}(p_l)$$
$$m_l = y_l \times \sigma_m + m_r$$

The correspondence sketch of a quantile and the respective mean is drawn in Figure 1.

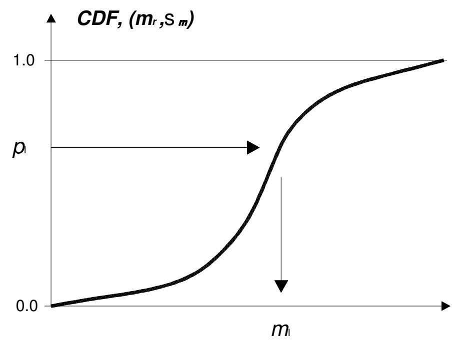


Figure 1: Sketch of the cumulative distribution function of distribution (m_r, σ_m), where $\sigma_m = S_m$.

Once the m_l values are calculated, the relation of these values and the mean of the original distribution (m_r) provide factors that are multiplied by each value of the original distribution. The factors (m_l / m_r) are ordered values because m_l corresponds to each quantile in the cdf. Then, the variable distributions have values that have the lower mean when l = 1 and the biggest mean when $l = L$.

$$z(\mathbf{u}_i^l) = z(\mathbf{u}_i) \times \frac{m_l}{m_r}, \quad i = 1,...,n$$

This factor is applied to strictly positive variables. Almost all data in the earth science are positive values including mineral grades, porosity and contaminant concentrations. The variance is not preserved in this transformation, but the most important statistic is the mean. The declustered weights of the variable distributions are the same as the original data because their spatial locations are the same.

3. Implementation

Simulation requires the original z-data to be transformed into y-values with a standard normal histogram, the normal score transform function can be derived through a graphical correspondence between the cumulative distribution of the original and standard normal variables (Goovaerts, 1997). The transformation process often uses the fixed ccdf of the original data; however, a different reference ccdf could be used. A simple way to transfer uncertainty in the mean through simulation is to use different reference distributions. An increase in global and local uncertainty is expected. A simple scenario explains the methodology, where a spherical semivariogram model is assumed, the distance between the $\mathbf{u}_{\square}$ location to be simulated and the sample $z(\mathbf{u}_1)$ corresponds to half of the range of the semivariogram (a/2). Since there is only one conditioning data, the conditional mean and variance is simplified to:

$$m_{c_{\square}} = \rho(\mathbf{h}) \cdot y(\mathbf{u}_1)$$
$$\sigma_{c_{\square}} = 1 - \rho(\mathbf{h})^2$$

Two scenarios are evaluated; the first scenario considers a fixed global distribution. The simulation uses the parameters of the original fixed global distribution to standardize its datum of 2.5 original units into normal scores. Then, the conditional mean kriging (0.313) and conditional variance kriging (0.902) are predicted for the unsampled location. An independent residual that follows a normal distribution with mean of zero and the conditional variance is drawn with classical Monte Carlo simulation. The simulated value is the addition of the conditional mean kriging and the residual for that location (Deutsch, 2002). Figure 2 shows the result of this simulation, where the output distribution for the unsampled location is illustrated.

The second scenario account for the simulation that uses a different reference distribution for the transformation of original values into normal scores and vice versa. The uncertainty in the mean of the input univariate distribution has a standard deviation of 0.2. The sampled location $z(\mathbf{u}_1)$ will have different ccdfs to transform to normal scores; $y(\mathbf{u}_1)$ in normal score unit takes values from 0.66 to 1.34. The product of those values and the weight kriging (0.313) gives different mean values and a constant variance kriging (0.902). Those values are sampled many times and back transformed into original units. The back transformation should use their respective transformation table matching the forward transform. Figure 3 shows the result of this simulation, where the output distribution for the $\mathbf{u}_{\square}$ unsampled location is wider than the previous simulation with fixed distribution.

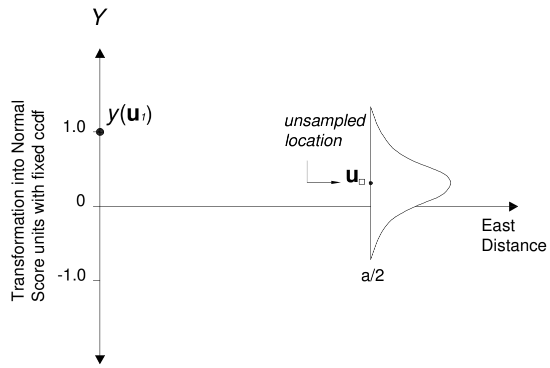


Figure 2: Sketch of simulation in one node using fixed ccdf $(1,1.5^2)$ like input parameter.

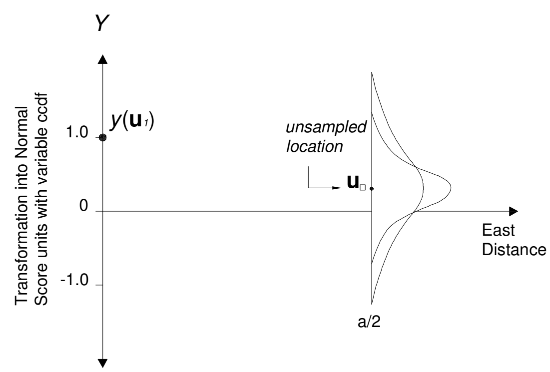


Figure 3: Sketch of simulation at one node using variable ccdf to transfer uncertainty in the original distribution to simulation.

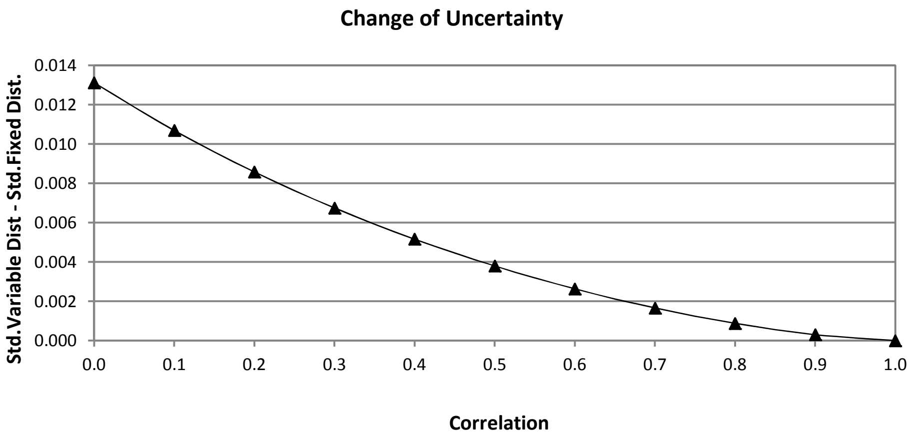
The number of reference distributions is denoted with the letter *L* and the number of realizations of every reference distribution is denoted with the letter *K.* The resulting mean and variance:

$$m_z = \frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K} Z_k^l(\mathbf{u}_{\square}) = 1.47 \qquad \sigma^2_{z_{unc}} = \frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K}\left(Z_k^l(\mathbf{u}_{\square}) - m_z\right)^2$$

As expected, the distribution of the output mean using uncertainty in the input parameter is wider than using a fixed input parameter. Uncertainty in the sampled location with fixed reference distribution is 1.41 and with variable reference distributions is 1.55.

4. Sensitivity Analysis

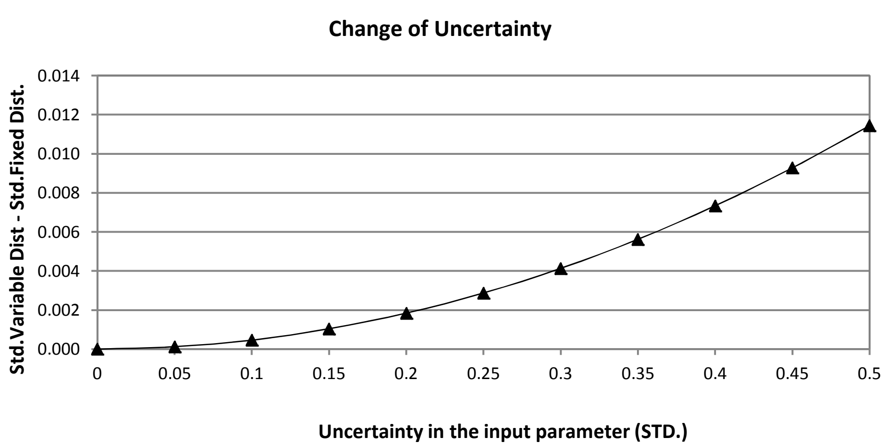
One thousand simulations are performed with parameter uncertainty. Each simulation considers variable transforms into normal scores. One thousand quantiles are used to draw the residuals. The change of the correlation between the conditioning data and the unsampled location are evaluated. As expected, the results show that the uncertainty goes down when the correlation increases. Figure 4 shows the less increase in uncertainty as the spatial correlation increases.



Data mean	1.00
Data STD	1.50
Uncertainty in mean of Input Distribution (STD.)	0.20
Conditioning value	1.00

Figure 4: Sensitivity analysis of the uncertainty with respect to the change of correlation, 1000 realizations are generated with fixed ccdf and with 100 variable ccdfs, the table shows the parameter used in the simulation where global mean and conditioning value are in original units.

The simulation with variable ccdfs and zero range of correlation show uncertainty of the node equal to 1.512; the simulation with fixed mean gives an uncertainty of 1.499. Moreover the change of the conditioning sample value does not change the uncertainty when the simulation uses a fixed distribution. This is expected because, under a Gaussian model, errors are independent of the data values and dependent only on the data configuration (Goovaerts, 1997); however, a change of uncertainty in the node is observed when a lognormal distribution is used. As expected, the uncertainty at the unsampled location increases as the uncertainty in the input parameter increases. Figure 5 shows the increase in uncertainty at the unsampled location as the input parameter uncertainty increases.



Data mean	1.00
Data STD	1.50
Conditioning value	1.00
Correlation	0.68

Figure 5: Sensitivity analysis of the uncertainty with respect to the change of UMID, 1000 realizations are generated with fixed cdf and with 100 variable cdfs, the table shows the parameters used in the simulation where global mean and conditioning value are in original units.

The same scenario of one conditioning sample is expanded to a grid of five nodes in the east direction and five nodes in the north direction. The size of the nodes is one unit. The change of the local uncertainty and global uncertainty is evaluated for this scenario. The uncertainty in the mean of the univariate distribution is accounted for in the generation of 1000 variables means. The sequential Gaussian simulation approach is applied. The conditioning data is located in the center of the domain; an exponential variogram model is used with range of 7.779 units because the covariance between the conditioning data and the closest node was set to 0.68. Figure 6 shows the change of the distribution of the global means when uncertainty in the mean of the distribution is incorporated to the process of simulation.

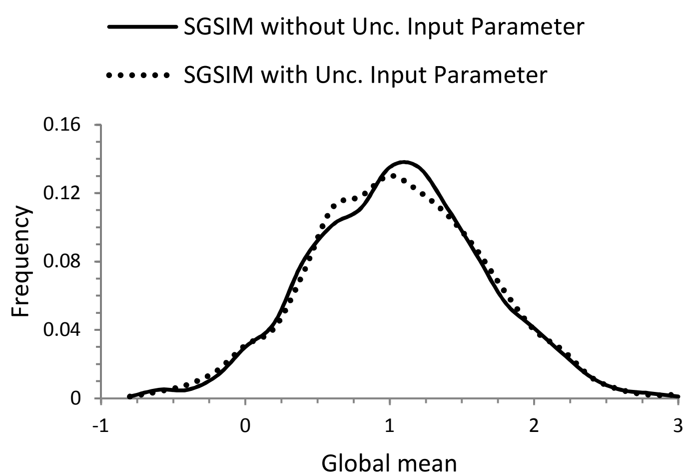


Figure 6: Spatial location of the conditioning sample $z(\mathbf{u})$ in the domain, where the covariance $z(\mathbf{u})$ to the nearest node is 0.688. The distribution of global means from SGSIM that use parameter uncertainty is compared with the one without parameter uncertainty.

Twenty five nodes are evaluated in a 2D map. Just as in the case of one node, two scenarios are evaluated; both of them run with the same random number seed. The increase of global uncertainty (std.) is from 0.60 to 0.61. Also, the increase in uncertainty at all the nodes is observed when the distribution of the residuals is drawn with quantiles instead of random numbers. Conversely, the sampling of the residuals with random numbers shows two nodes with a very small reduction of local uncertainty.

5. Practical Considerations

Many techniques to evaluate uncertainty in the input parameters are available. All of them give reasonable output; however, it is important to keep in mind that some scenarios or phases of a project development require more parameters than other. Parameters like spatial correlation and finite domain must be taken account when the project has enough data to define the domain.

The input distribution should be representative of the domain or volume to be evaluated. The limits in the tails of the distribution should be carefully defined. A wrong definition of the tails value could generate some artifact in the transformation of the values into Gaussian units. The tail values may need to be chosen separately for each variable ccdf.

The red data file is available in the CCG network and has 68 samples through a vein. There are samples of gold, silver, copper and zinc. The thickness of the samples is between 0.13 and 18.86 meters. The spatial distance between the samples is about 30 meters. The gold value is evaluated. A semivariogram model of the gold values in normal score units is required for the simulation-based approaches.

$$\gamma(\mathbf{h}) = 0.44 Exp_{\substack{ah1=100\\ ah2=90}}(\mathbf{h}) + 0.56 Exp_{\substack{ah1=250\\ ah2=95}}(\mathbf{h})$$

Two structures were required to model the experimental variogram, which *ah1* is the mayor range in the 15° azimuth and *ah2* is the minor range in perpendicular direction to *ah1.* The evaluations are done on a domain of size 500 meters × 600 meters using a discretization of 5 meters × 5 meters blocks. The mean of the input distribution is 1.415 ppm with a standard deviation of 1.288. The uncertainty in the mean of the input distribution is the standard deviation of 0.360 using conditional finite domain.

One hundred variable cdfs are created with a *genrefdist.for* program that was developed for this approach. One hundred sets of simulations are performed with their respective variable ccdf. Those ccdfs are used as reference ditribution in the program *sgsim.for*. The output files are gathered with a program *mixsim.for*. The same number of simulations is executed for the fixed ccdf. The two sets of simulations are compared and the increase in uncertainty is illustrated in a 2D map. The increase of the local uncertainty is visible in zones where the samples show high spatial variability. For instance, from the Figure 7 the samples that have 0.001 ppm the lower quantity of gold are neighbours with samples with thousand times more high values.

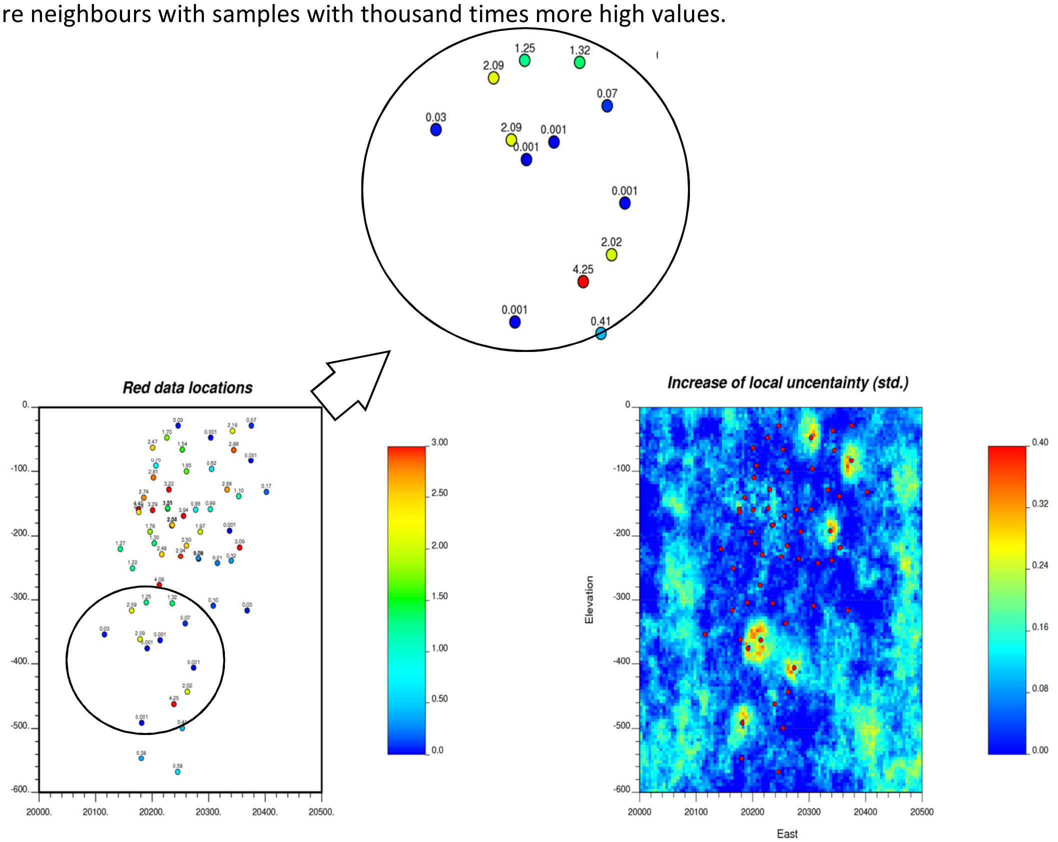


Figure 7: Location of the red data and map of increase in local uncertainty because the uncertainty in the input parameter is transferred to the simulation.

Besides those zones, the other zones that present a considerable increase in uncertainty is in zones that are located far from the conditioning data.

The local uncertainty at the nodes using a fixed reference distribution is compared to the one using different reference distributions by scatter graphic. Positive correlation is observed in Figure 8.

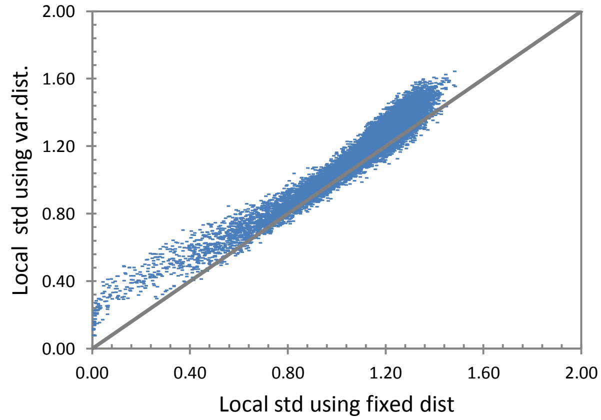


Figure 8: Increase in uncertainty at each node after being simulated with different reference distributions.

Every realization or map gives a unique mean that change through the realizations. The standard deviation of these means is defined as a global uncertainty. The global uncertainty with fixed ccdf is 0.12 and a narrow shape of the global means is given between 1.03 ppm and 1.64 ppm. The second scenario when the simulation take account the uncertainty in the mean of the input univariate distribution, the deviation of the global means increase to 0.24, that is, the tails of the distributions of the global means is expanded from 0.77 ppm to 2.30 ppm. Figure 9 shows the increase in uncertainty when is transferred the uncertainty of the input parameter to the simulation.

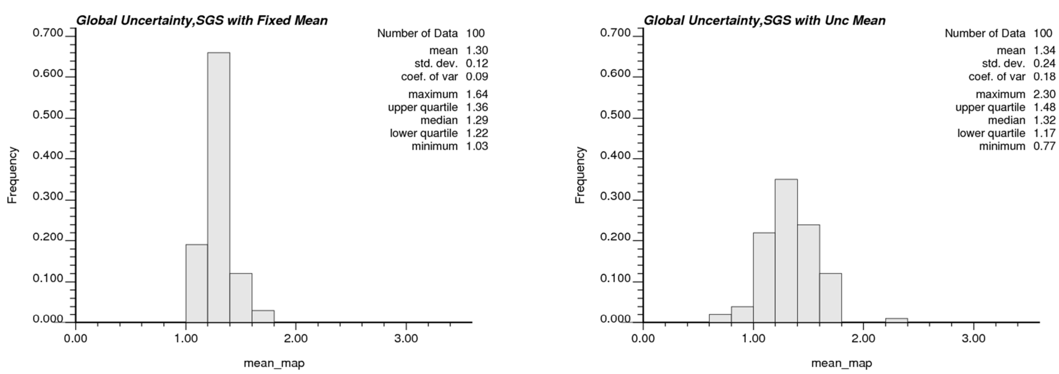


Figure 9: Change of global uncertainty, fixed ccdf right histogram and variable ccdfs left histogram.

The example shows that an increase in uncertainty is observed in local scale or zones close to the data and in long scale or zones that are far from the data. Also, the uncertainty at the locations of the data is zero with fixed and variable ccdf. The narrow uncertainty of the global mean is the optimal scenario provided that this uncertainty is accurate; however, uncertainty in the input parameter should not be ignored in the simulation process.

6. Conclusions

An important contribution of this paper is demonstrating how uncertainty in the mean of the input distribution is transferred through geostatistical simulation for a more complete understanding of uncertainty. Geostatistical simulation is usually performed with a fixed input distribution; this fixed univariate distribution comes from the input data and assumes a mean without uncertainty.

Multiple distributions could be constructed and used in geostatistical simulation as reference distributions. Original values are transformed into Gaussian units according to a specified reference distribution. The uncertainty in the mean of the univariate distribution is accounted by changing the reference distribution for transformation.

Any Gaussian algorithm for simulation could be used. A change in local and global uncertainty is expected when the uncertainty is transferred through the simulation process.

7. References

Arik, A. (1999). An Alternative Approach to resource Classification. *APCOM Proceeding, Computer Applications in the Mineral Industries*, 45-33.

Deutsch, C. V. (2000). *Special Topics in Geostatistics.* Alberta: University of Alberta.

Deutsch, C. V. (2002). *Geostatistical Reservoir Modeling.* New York: Oxford University Press, Inc.

Deutsch, C. V. (2004). A statistical Resampling Program for Correlated Data, Spatial Boostrap. *Six Annual Report of the Centre for Computational Geostatistics.* Edmonton: Department of Civil & Environmental Engineering University of Alberta.

Deutsch, C. V., & Journel, A. (1998). *GSLIB: Geostatistical Softare Library and User's Guide* (2nd Edition ed.). New York: Oxford University Press.

Deutsch, C. V., Leuangthong, O., & Ortiz C., J. (2006). A Case for Geometric Criteria in Resources and Reserves Classification. *Eight Annual Report of the Centre for Computational Geostatistics* (p. 21). Edmonton: Department of Civil & Environmental Engineering-University of Alberta; Department of Mining Engineering-University of Chile.

Dominy, S. C., Noppé, M. A., & Annels, A. E. (2002, October). Errors and Uncertainty in Mineral Resource and Ore Reserve Estimation: The Importance of Getting it Right. *Exploration and Mining Geology, 11*, 77-98.

Emery, X. (2008). Uncertainty modeling and spatial prediction by multi-Gaussian kriging: accounting for an unknown mean value. *Computer & Geosciences 34*, 1431-1442.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* New York: Oxford University Press.

Journel, A. (2004). Roadblocks to the Evaluation of Ore Reserve, The simulation Overpass and Putting More Geology into Numerical Model of Deposits. In R. Dimitrakopoulos (Ed.), *Ore Modeling and Strategic Mine Planning - Uncertainty and Risk Management Models* (pp. 17-20). Perth: AusIMM.

Journel, A. G. (1994). Resampling from Stochastic Simulation. *Environmental and Ecological Statistics, 1*, 63-91.

Leuangthong, O., Khan, K. D., & Deutsch, C. V. (2008). *Solved Problem in Geostatistics.* Hoboken, New Jersey.: John Wiley & Sons, Inc.

Leuangthong, O., McLennan, J., & Deutsch, C. V. (2005). *Acceptable Ergodic Fluctuations and Simulation of Skewed Distributions.* Edmonton: Centre for Computational Geostatistics, Department of Civil&Environmental Engineering University of Alberta.

Leuangthong, O., McLennan, J., & Deutsch, C. V. (2005). Acceptable Ergodic Fluctuations and Simulation of Skewed Distributions. *Seventh Annual Report of the Centre for Computational Geostatistics.* Edmonton: Department of Civil&Environmental Engineering-University of Alberta.

Pyrcz, M., & Deutsch, C. (2002). Two artifacts of probability field simulation. *Mathematical Geology, 33*, 775-800.

Pyrcz, M., & Deutsch, C. (2002). Two atifacts of probability field simulation. *Mathematical Geology, 33*, 775-800.

Rose, P. R. (2001). *Risk Analysis and Management of Petroleum Exploration Ventures.* Tulsa, Oklahoma: The American Association of Petroleum Geologist.

Strunk Jr., W., & White, E. (2009). *The Elements of Style* (Fiftieth Anniversary ed.). New York, United States: Pearson Education, Inc.