On the Scaling and Use of Multivariate Distributions in Geostatistical Simulation

Oy Leuangthong, Julián M. Ortiz and Clayton V. Deutsch

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

Geostatistical site characterization often requires the simultaneous modeling of numerous correlated variables. Reproducing complex features such as non linearity, stoichiometric constraints and heteroscedasticity is difficult. The multivariate probability distribution must be discretized and used directly in the construction of numerical models. In presence of clustered and/or biased data, it is necessary to scale these multivariate distributions to representative univariate distributions. A multivariate scaling approach is developed that preserves multivariate features whilst imposing representative univariate distributions.

Most geostatistical techniques do not permit the direct input of a multivariate distribution. The multivariate Gaussian model is too simple; it does not permit use of realistic multivariate distributions. A stepwise transformation approach could be used as a pre-processing step to establish variables that are univariate Gaussian. The stepwise approach requires an equal sampling of the different variables and samples of the same scale; an alternative must be considered when these conditions are not met. The direct approach, based on the simple cokriging principle and covariances, could be combined with the multivariate distribution scaling approach for cosimulation. These alternatives are developed and guidelines presented for their application.

Introduction

Geostatistical simulation is used in several areas of the earth sciences to quantify uncertainty in mineral resources (Journel, 1974), assess the risk of exceeding critical thresholds in contamination studies (Kyriakidis and Journel, 2001), provide multiple possible realizations of the petrophysical properties of a petroleum reservoir for flow simulation (Deutsch, 2002), etc. In most of these cases, multiple variables have been measured at sample locations and should be used jointly to improve the prediction of the variables at unsampled locations. Furthermore, these variables may have different volumetric supports. Integration of information from multiple sources requires the use of multivariate techniques (Wackernagel, 2003). However, most of these techniques rely on some Gaussian assumption that makes it difficult to handle complex relationships such as non linearity, stoichiometric constraints and heteroscedasticity. A Gaussian alternative that allows reproducing these features is the stepwise conditional transformation (Leuangthong and Deutsch, 2003); however, integration of data at different support remains a limitation for this method.

This paper discusses some of the issues encountered when dealing with complex multivariate distributions that cannot be correctly handled by conventional multivariate techniques. The direct approach to simulate random variables is presented. Since in this framework it is possible to determine the conditional distributions for each variable, there is a need to update the multivariate distribution, honouring both the representative univariate distributions and the characteristic features of the multivariate relationships. An iterative approach is proposed and illustrated for a

simple example; the approach could be implemented in a sequential direct cosimulation framework to handle simulation of complex multivariate relationships.

Direct approach to multivariate simulation

Direct simulation, that is, simulating a variable without having to work with a transformation of it, has been an important research area in geostatistics (Caers, 2000; Journel, 1993; Oz and others, 2003; Soares, 2001).

The advantage of considering a direct approach is that multiscale data can be easily integrated in simulation. Transformation to normal scores impedes the use of block data or the direct simulation at a different support from the sample data, because averages of the transformed data do not translate into the correct averages in the original variables. Working in direct space allows immediate integration and use of data at different scale, without biasing the averages when a simulation at a different support is performed.

The paradigm of direct simulation relies on the fact that, as long as the simulation is done sequentially, covariance reproduction is ensured when the simulated values are generated from a conditional distribution whose mean and variance are calculated by simple kriging from the sample data and previously simulated nodes (Journel, 1993). The main question and one of the problems of direct sequential simulation is the reproduction of the global histogram: although any shape is allowed in the conditional distributions for covariance reproduction, these shapes determine the final histogram of the realizations. It is necessary to have an approach for obtaining a set of conditional distributions that will return the right histogram after sequentially drawing from them. A solution consists on using the Gaussian transform instrumentally to get a set of conditional distribution shapes, parameterized by their conditional means and variances (Oz and others, 2003). The procedure consists of building a lookup table where a mean and variance in original units is linked to a conditional distribution shape. The conditional distributions in original units are obtained by back-transforming the quantiles of a non-standard Gaussian distribution, using the link between the global distribution and a standard Gaussian distribution (the conventional normal score transformation table). The back-transformed quantiles of the non-standard Gaussian distribution can be numerically integrated to get the mean and variance of the distribution in original units, hence that conditional distribution is parameterized by these two values.

Direct sequential simulation proceeds just as sequential Gaussian simulation. The nodes are visited randomly and at every node the mean and variance of the conditional distribution in original units are calculated by simple kriging of the data and previously simulated nodes. Then, a conditional distribution is retrieved from the lookup table and used to draw the simulated value that will condition all subsequently simulated values. This method has proven to perform well (Oz and others, 2003) hence this principle can be extended to multivariate simulation. Multivariate simulation requires the knowledge of the multivariate conditional distribution. This distribution is characterized by the vector of conditional means, the conditional variance-covariance matrix, and its shape.

Conditional means and variance-covariance matrix can be obtained by solving a cokriging system. Consider a set of variables $Y_1,...,Y_P$ and assume, without loss of generality, that their means are zero. These variables can have any volumetric support denoted as $\left\{V_p(\mathbf{u}_\alpha^p);\alpha=1,...,n_p;p=1,...,P\right\}$ and are centered at locations $\left\{\mathbf{u}_\alpha^p;\alpha=1,...,n_p;p=1,...,P\right\}$. We can

write the cokriging estimate (the conditional mean) and the estimation variance at a location $\mathbf{u}$ and with a support $V_i(\mathbf{u})$ as:

$$\left[Y_i(\mathbf{u})\right]^*_{SCoK} = \sum_{p=1}^{P}\sum_{\alpha=1}^{n_p} \lambda_\alpha^p \, Y_p(\mathbf{u}_\alpha^p)$$

$$\sigma_E^2(\mathbf{u}) = \overline{C}(V_i(\mathbf{u}), V_i(\mathbf{u})) + \sum_{p=1}^{P}\sum_{q=1}^{P}\sum_{\alpha=1}^{n_p}\sum_{\beta=1}^{n_q} \lambda_\alpha^p \, \lambda_\beta^q \, \overline{C}(V_p(\mathbf{u}_\alpha^p), V_q(\mathbf{u}_\beta^q)) - 2\sum_{p=1}^{P}\sum_{\alpha=1}^{n_p} \lambda_\alpha^p \, \overline{C}(V_i(\mathbf{u}), V_p(\mathbf{u}_\alpha^p))$$

with $\overline{C}(V_p(\mathbf{u}_\alpha^p), V_q(\mathbf{u}_\beta^q)) = \dfrac{1}{V_p \cdot V_q} \int_{V_p(\mathbf{u}_\alpha^p)} \int_{V_q(\mathbf{u}_\beta^q)} C(v-w)\, dv\, dw$ the average covariance with the head of the separation vector spanning $V_p(\mathbf{u}_\alpha^p)$ and its tail spanning $V_q(\mathbf{u}_\beta^q)$.

The mean and variance are calculated by solving the following system of equations:

$$\sum_{q=1}^{P}\sum_{\beta=1}^{n_q} \lambda_\beta^q \, \overline{C}(V_p(\mathbf{u}_\alpha^p), V_q(\mathbf{u}_\beta^q)) = \overline{C}(V_p(\mathbf{u}_\alpha^p), V_i(\mathbf{u})) \qquad \forall p = 1,...,P;\ \forall \alpha = 1,...,n_p$$

Cokriging provides the parameters to obtain the univariate conditional distributions for all variables using the lookup table previously described, however, the entire conditional multivariate distribution cannot be retrieved. The next section presents an approach to scale the global multivariate distribution in order to honour the univariate conditional distributions and the characteristics of the multivariate relationship.

Note that Gaussian methods impose a multivariate behaviour which is linear and homoscedastic, which may be deemed inappropriate. Furthermore, Gaussian simulation does not allow integrating block data or directly simulating at block support.

A multivariate scaling approach

Honouring the multivariate features is a key aspect to predict the risk when multiple variables are relevant in a response variable. The shape of the multivariate global distribution is used as a reference for characterizing the shapes of the conditional multivariate distributions. The iterative approach proposed in this section takes the global multivariate distribution and iteratively scales it. Initially, the scaling is done by using the product of the ratios between the probability densities of the global univariate distributions to the density of the conditional univariate distributions. Subsequently, the ratio considers the updated univariate distribution, which is computed by integrating the updated multivariate distribution obtained after this iteration, and the target conditional univariate distribution. The updating is repeated until a good match of all relevant statistics is achieved: means and variances of the univariate conditional distributions, and correlations between pairs of variables.

Let $f^{(0)}_{Y_1,...,Y_P}(y_1,...,y_P)$ be the global multivariate distribution, $f^{(0)}_{Y_i}(y_i), \forall i = 1,...,P$ the global univariate distributions, and $f^{(k)}_{Y_i}(y_i), \forall i = 1,...,P$ the univariate conditional distributions. The updating technique can be summarized in the following steps for a P-variate distribution:

1. Initialize $f^{(k)}_{Y_i}(y_i), \forall i = 1,...,P$ to the target univariate conditional distribution, $\forall k$ iterations. Set $k = 1$.

2. Update the multivariate global distribution by the ratios of the univariate conditional distributions to the current univariate distributions, that is:

$$f^{(k)}_{Y_1,...,Y_P}(y_1,...,y_P) = f^{(k-1)}_{Y_1,...,Y_P}(y_1,...,y_P) \cdot \prod_{i=1}^{P} \frac{f^{(k)}_{Y_i}(y_i)}{f^{(k-1)}_{Y_i}(y_i)}$$

3. Obtain the updated univariate distributions by integrating the multivariate conditional distribution:

$$f^{(k)}_{Y_i}(y_i) = \int ... \int f^{(k)}_{Y_1,...,Y_P}(y_1,...,y_P)\, dy_1 ... dy_{i-1}\, dy_{i+1} ... d_P \qquad \forall i = 1,...,P$$

4. Calculate summary statistics: means and variances of the updated univariate distributions and correlation coefficients between variables of the updated multivariate distribution.

5. Check if these statistics match the parameters of the univariate conditional distributions and conditional multivariate distribution within some tolerance. If this condition is matched, then stop; otherwise, set $k = k+1$ and goto 2.

Example

A simple example is presented to show the performance of the scaling approach for a non-linear relationship between Ni and Fe from a nickel laterite deposit. This is illustrated using the bivariate frequency plot shown in Figure 1. The global bivariate frequency resulting from the multivariate scaling approach is also shown in Figure 1. A visual comparison of the bivariate frequencies shows good reproduction of the bivariate features, with the reference correlation of 0.091 and the updated correlation coefficient of 0.088. A comparison of the reference and the updated univariate cumulative distribution function for Ni and Fe shows virtually exact reproduction of the univariate statistics.

Implementing direct sequential cosimulation of multiscale data

Implementing this scaling approach in the direct sequential simulation framework for multiple variables measured at different supports calls for (1) the cokriging formalism for multiscale data, (2) the determination of the conditional distributions linked only to their conditional means and variances through the lookup table as proposed by Oz and others (2003), (3) the use of the iterative scaling approach proposed in the previous section, and (4) drawing from the multivariate distribution by Monte-Carlo simulation one variable at a time with increasing levels of conditioning. The main steps are summarized below:

1. Define a random path to visit the uninformed locations to simulate. These do not have to be necessarily at point-support.

2. At each location in the path:

    a. Perform a search of all nearby original data of any variable and at any scale and previously simulated locations.

    b. Determine the conditional means and variances of each variable by a simultaneous cokriging of all variables.

    c. Obtain the univariate conditional distribution of each variable.

d. Obtain the multivariate conditional distribution using the iterative scaling approach proposed, honouring the univariate conditional distributions and the multivariate relationships found in the global multivariate distribution.

e. Draw a simulated vector from the multivariate conditional distribution as follows:

    i. Draw a simulated value for one variable from one of the univariate conditional distributions.

    ii. Draw a simulated value for another variable using the conditional distribution given the value drawn for the first variable.

    iii. Repeat increasing the level of conditioning until the full vector has been simulated.

f. Visit the next location.

Conclusions

An approach to scale a multivariate distribution is presented that allows characterizing conditional distributions that honour the conditional univariate distributions and the relationship between the variables as depicted in the global multivariate distribution. The approach works iteratively, by modifying the global multivariate distribution with the product of ratios between the univariate probability densities of the conditional univariate distribution to the global univariate distribution. Once repeated enough times, this updating technique leads to multivariate distributions whose marginals are the correct univariate conditionals and the correlation and features between multiple variables are reproduced. The scaling algorithm could be used in a direct sequential simulation framework to allow simulation of multivariate multiscale data. An integrated approach has been presented, where the direct paradigm is presented in all generality. This approach promises to be a powerful alternative to the limitations of multivariate Gaussian simulation.

Acknowledgements

The authors acknowledge the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta and the Chair in Ore Reserve Evaluation at the University of Chile sponsored by Codelco Chile.

References

Caers, J., 2000: Adding Local Accuracy to Direct Sequential Simulation. Mathematical Geology, 32(7): 815-850.

Deutsch, C. V., 2002: Geostatistical Reservoir Modeling, Oxford University Press, pp. 376.

Journel, A. G., 1974: Geostatistics for conditional simulation of orebodies, Economic Geology, 69: 673-687.

Journel, A. G., 1993: Modeling Uncertainty: Some Conceptual Thoughts, in Dimitrakopoulos, R., ed., Geostatistics for the Next Century, Kluwer, Dordrecht, Holland, p. 30-43.

Kyriakidis, P. C., and Journel, A. G., 2001: Stochastic modeling of atmospheric pollution: a spatial time-series framework. Part II: application to monitoring monthly sulfate deposition over Europe, Atmospheric Environment, 35(13): 2339-2348.

Leuangthong, O., and Deutsch, C. V., 2003: Stepwise Conditional Transformation for Simulation of Multiple Variables, Mathematical Geology, 35(2): 155-173.

Oz, B., Deutsch, C. V., Tran, T. T., and Xie, Y., 2003: DSSIM-HR: A Fortran 90 Program for Direct Sequential Simulation with Histogram Reproduction, Computers & Geosciences, 29(1): 39-51.

Soares, A., 2001: Direct Sequential Simulation and Cosimulation, Mathematical Geology, 33(8): 911-926.

Wackernagel, H., 2003. Multivariate geostatistics: an introduction with applications, 3rd edition, Springer, Berlin, 387 p.

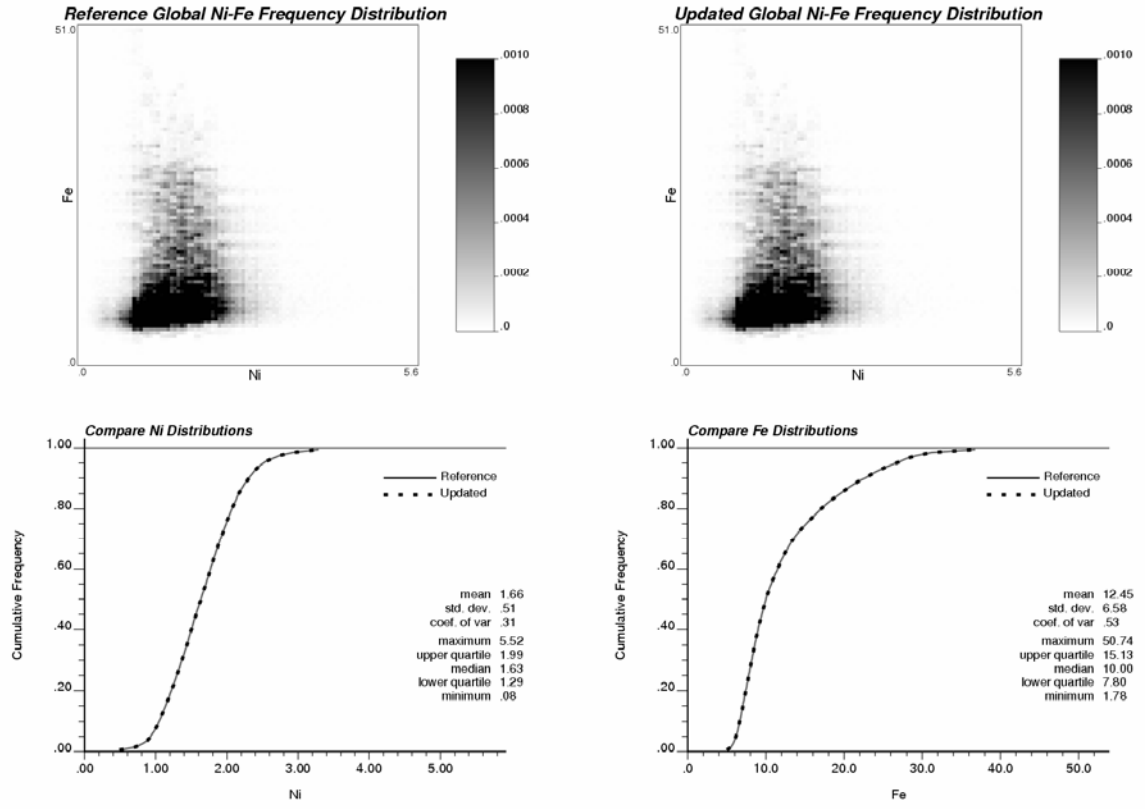

Figure 1: Multivariate scaling using a nickel laterite data set: comparison of reference vs. updated global bivariate distribution (top row) and univariate Ni and Fe distributions (bottom row).