Short Note:
Bias in SIS due to order relation corrections and a quick fix

Julián M. Ortiz (jmo1@ualberta.ca)
Department of Civil & Environmental Engineering
University of Alberta

Abstract

The conventional correction of order relations in sequential indicator simulation introduces a bias for extreme thresholds; the resulting cdf values are closer to the median. A dynamic order relations correction is proposed to make the estimator unbiased after order relation corrections.

Sequential Indicator Simulation

Sequential indicator simulation allows the characterization of a regionalized variable by a set of thresholds $z_1, ..., z_K$. The sample data are coded as indicators at every threshold and the conditional probability at unsampled locations is estimated by simple indicator kriging. The stationary simple kriging estimate of the indicator at a given threshold is written:

$$\begin{aligned} [i(\mathbf{u}; z_k)]^*_{SK} &= [Prob\{Z(\mathbf{u}) \leq z_k | (n(\mathbf{u}))\}]^*_{SK} \\ &= \textstyle\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{SK}(\mathbf{u}; z_k) \cdot i(\mathbf{u}_\alpha; z_k) + [1 - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{SK}(\mathbf{u}; z_k)] \cdot F(z_k) \end{aligned}$$

where the weights $\lambda_\alpha^{SK}(\mathbf{u}; z_k)$ are the unique solution of the simple kriging system:

$$\textstyle\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta^{SK}(\mathbf{u}; z_k) \cdot C_I(\mathbf{u}_\beta - \mathbf{u}_\alpha; z_k) = C_I(\mathbf{u} - \mathbf{u}_\alpha; z_k) \quad \alpha = 1, ..., n(\mathbf{u})$$

Notice that a stationary indicator covariance function $C_I(\mathbf{h}; z_k)$ must be inferred for each threshold.

Order Relation Deviations

The estimated probabilities $[i(\mathbf{u}; z_k)]^*_{SK}$, $k = 1, ..., K$ generated through indicator kriging must satisfy the conditions of a cumulative distribution: they have to be non-decreasing between 0 and 1 [1, 2, 3, 4].

The kriged indicator value can lie outside the interval [0,1] because the kriged estimate may be a non-convex linear combination of the conditioning data. Lack of data in some classes and differences in the variogram models from one threshold to the next are important factors that may produce a non-increasing function [2, 5].

The *a posteriori* upward and downward correction of the conditional cumulative distribution function works well in general, as documented by Deutsch and Journel [2] (**Figure 1**). Another more difficult solution is to constrain the kriging system, so that it satisfies the order relations by construction [3].

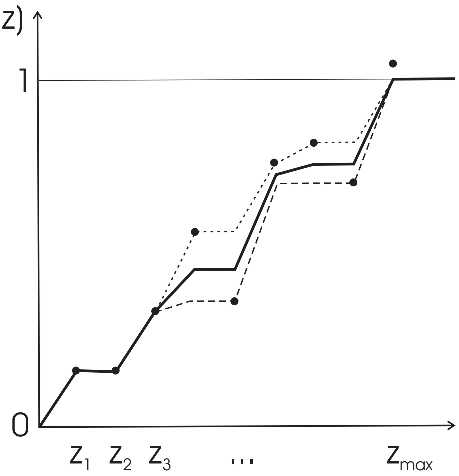

Figure 1: Upward and downward correction for order relation deviations. The corrected cumulative conditional distribution is the thick line.

Bias in the Estimator

The kriging estimate $[i(\mathbf{u};z_k)]^*_{SK}$ is unbiased. However, since corrections are required to ensure that a valid conditional distribution is built prior to simulating a value, the estimator is no longer unbiased.

The corrected estimator can be written:

$$[i(\mathbf{u};z_k)]^{corrected}_{SK} = [i(\mathbf{u};z_k)]^*_{SK} + \Delta_{upward}(z_k) - \Delta_{downward}(z_k)$$

where the expected value of the corrected estimate is no longer unbiased:

$$E\{[I(\mathbf{u};z_k)]^{corrected}_{SK}\} \neq E\{[I(\mathbf{u};z_k)]^*_{SK}\}$$

because for high thresholds, such that $F(z_k) > 0.5$, $E\{\Delta_{downward}(z_k)\} > E\{\Delta_{upward}(z_k)\}$, and for low thresholds with $F(z_k) < 0.5$, $E\{\Delta_{downward}(z_k)\} < E\{\Delta_{upward}(z_k)\}$.

The bias introduced by order relation corrections depends on the threshold that is being estimated.

Considering a binary realization, that is, when only one threshold is being used, say the ninetieth percentile, the bias is introduced by correcting more often deviations due to having an estimate greater than one, than deviations due to the estimate being less than zero. Overall, the estimated values are no longer unbiased.

Figure 2 shows the histogram of corrections required during a run of sequential indicator simulation with a threshold at the 90^{th} percentile. Overall, corrections are biased, giving a non-zero average. Furthermore, when inspecting the histograms of positive and negative corrections, we note that positive corrections are made in more than 95% of the cases where a correction is required and the average of the positive corrections is much smaller than the average of the negative corrections. However, negative corrections are fewer than positive corrections, and despite their larger magnitude, they are not enough to counterbalance the positive corrections. This leaves an overall positive bias in the estimation of the probabilities. A similar problem can be seen when considering different thresholds. When the median is used, the corrections for values above one and below zero are similar, cancelling each other and generating no bias.

A Possible Fix

One way to overcome this problem is to dynamically correct for the bias introduced, every time a correction is made. This has been implemented with favorable results. The idea is to keep track of the last order relation correction made, and to adjust the next estimate by that amount, in order to generate overall unbiased realizations. Consider that the node at location $\mathbf{u}_0$ has been simulated. The simple indicator kriging estimate at that location is:

$$[i(\mathbf{u}_0; z_k)]^*_{SK} = \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha^{SK}(\mathbf{u}_0; z_k) \cdot i(\mathbf{u}_\alpha; z_k) + [1 - \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha^{SK}(\mathbf{u}_0; z_k)] \cdot F(z_k)$$

where $n(\mathbf{u}_0)$ is the number of sample data and previously simulated nodes found in a search neighborhood around $\mathbf{u}_0$.

If a correction is required for this node at the threshold z_k, then the corrected estimate is:

$$[i(\mathbf{u}_0; z_k)]^{corrected}_{SK} = \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha^{SK}(\mathbf{u}_0; z_k) \cdot i(\mathbf{u}_\alpha; z_k) + [1 - \sum_{\alpha=1}^{n(\mathbf{u}_0)} \lambda_\alpha^{SK}(\mathbf{u}_0; z_k)] \cdot F(z_k) + \Delta_0$$

The value Δ_0 can be positive or negative, depending if the correction increases or decreases the estimated cdf value. This value is kept in memory for the subsequent node simulated. This node is randomly picked among all uninformed nodes in the domain. Say the next node to simulate is located at $\mathbf{u}_1$. The dynamically corrected estimate is:

$$[i(\mathbf{u}_1; z_k)]^{**}_{SK} = \sum_{\alpha=1}^{n(\mathbf{u}_1)} \lambda_\alpha^{SK}(\mathbf{u}_1; z_k) \cdot i(\mathbf{u}_\alpha; z_k) + [1 - \sum_{\alpha=1}^{n(\mathbf{u}_1)} \lambda_\alpha^{SK}(\mathbf{u}_1; z_k)] \cdot F(z_k) + \Delta_0$$

where the superscript ** denotes the dynamically corrected estimate.

Again, this estimate may require correction for order relation deviations:

$$[i(\mathbf{u}_1; z_k)]^{corrected}_{SK} = [i(\mathbf{u}_1; z_k)]^{**}_{SK} + \Delta_1$$

The new correction factor Δ_1 is kept to correct the next node in the random path.

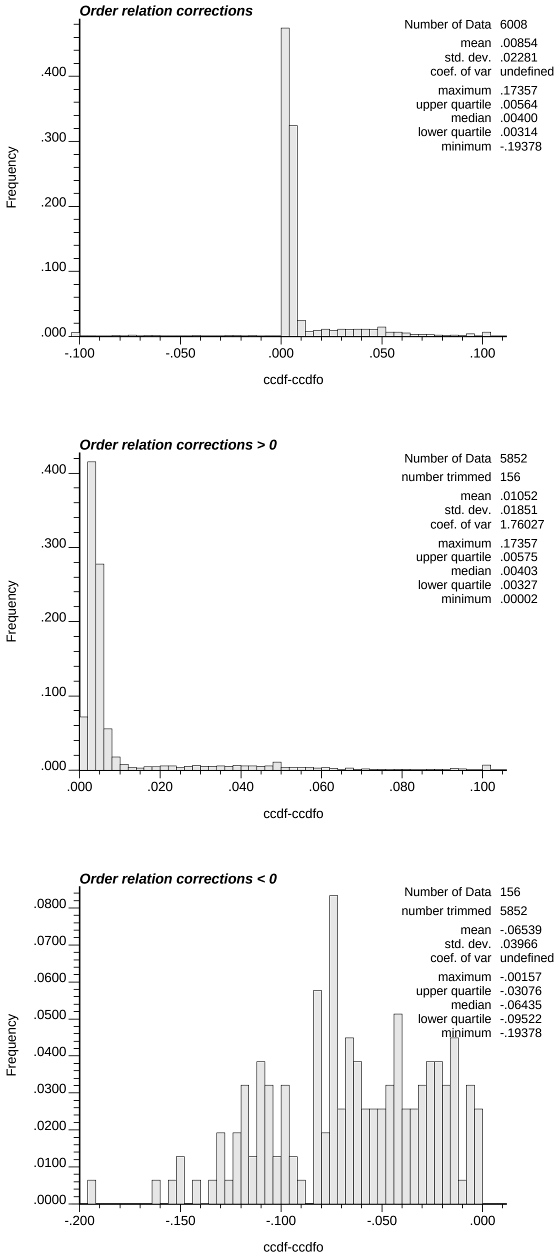

Figure 2: Histograms of order relation corrections in `SIS`. The top histogram shows all corrections together, the middle one shows the negative corrections and the bottom one shows the positive order relation corrections.

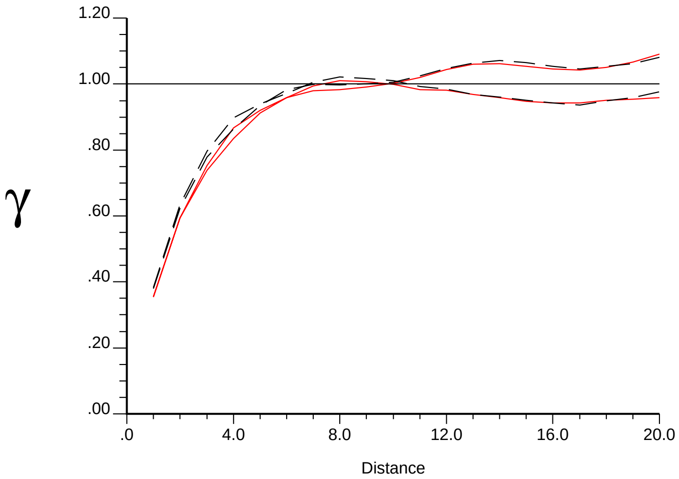

Figure 3: Experimental indicator variogram before (continuous line) and after dynamically correcting for the bias introduced by order relation corrections (dashed line) in `SIS`.

This dynamic correction generates a slight increase in the nugget effect, which is seen as a shift in the experimental variogram of the realization. The change is not significant if the corrections are small (**Figure 3**).

A histogram showing the distribution of output proportions from 100 realizations generated through sequential indicator simulation is presented in **Figure 4**. The histogram after correcting for the bias is also shown. The realizations were aimed at generating a proportion of 10% below the threshold. Notice the slight inflation in the variance of the distribution after the correction.

The magnitude of the order relation corrections will dictate if a dynamic correction for `SIS` is required. However, it is known that `SIS` performs well without that correction, if the following conditions are met:

- Enough conditioning information is available.
- The size of the simulated domain is large with respect to the range of correlation of the variogram.
- Multiple grid search is used to simulate.

Conclusions

Sequential indicator simulation generates a bias in the output proportions at extreme thresholds. A dynamic correction is proposed that shows satisfactory results. The correction consists on keeping track of the last correction for order relations deviations and correcting the estimate at the subsequent node in the random path.

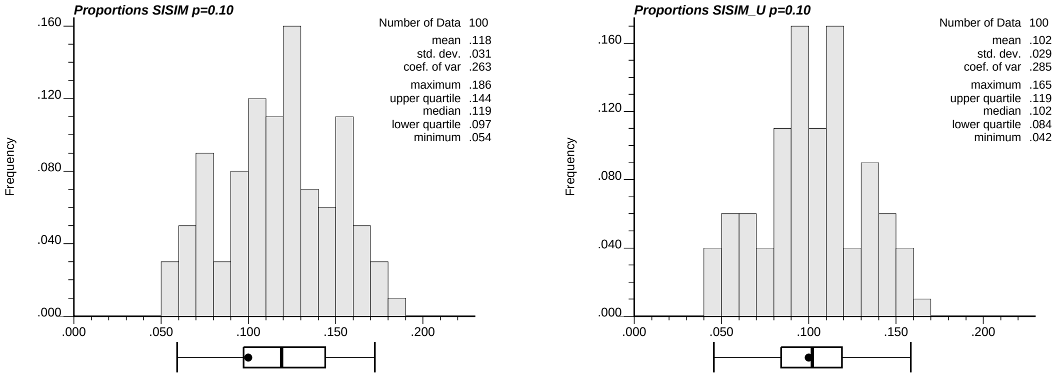

Figure 4: Histogram of output proportions before (left) and after (right) applying the dynamic correction in sequential indicator simulation.

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, 2nd edition, 1998.

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

[4] A. G. Journel. The place of non-parametric geostatistics. In G. Verly, M. David, A. G. Journel, and A. Marechal, editors, *Geostatistics for natural resources characterization*, volume 1, pages 307–335. Reidel, Dordrecht, Holland, 1984.

[5] A. G. Journel and D. Posa. Characteristic behavior and order relations for indicator variograms. *Mathematical Geology*, 22(8):1011–1025, 1990.