Follow Up on HISIM: Hierarchical Indicator Simulation

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

Abstract

The idea of simulating indicators hierarchically in order to avoid order relations and to set a framework suitable to incorporating multiple point statistics was proposed in last year's report. The implementation failed in that indicator variograms could not be reproduced for all thresholds. A loss in freedom from one threshold to the next was misinterpreted. The randomness was coming from random drawing of the nodes at high thresholds due to little difference between the probability of informed and uninformed nodes. In this note we explore several ways to fix this problem. A hierarchical implementation of sequential indicator simulation (SIS), along with methods that combine the SIS paradigm and the hierarchical idea, are also presented. Although some of the techniques give reasonable results, the problem remains largely unsolved from a theoretical point of view.

The original idea

The idea was to simulate one threshold at a time starting at the highest [3]. This can be seen as an erosion algorithm, where all nodes start higher than the highest cutoff, and then they are pushed down based on their probabilities of being below each threshold.

At a given threshold z_k, the conditioning data are coded as indicators:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1, & \text{if } z(\mathbf{u}_\alpha) \leq z_k \\ 0, & \text{otherwise} \end{cases} \quad k = 1, ..., K$$

where $z(\mathbf{u}_\alpha)$ is the value at the data location $\mathbf{u}_\alpha$.

The idea is to calculate the probability of every node being lower than the current threshold. This is done by simple kriging the indicators. The mean is the correct proportion from the global distribution.

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

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

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

Notice that a covariance indicator function $C_I(\mathbf{u} - \mathbf{u}_\alpha; z_k)$ (or, assuming stationarity, $C_I(\mathbf{h}; z_k)$), has to be inferred for each threshold.

Once the probabilities are known for every node, a node is chosen by Monte Carlo simulation, that is, a uniform random number between zero and one is drawn and the nodes are visited in

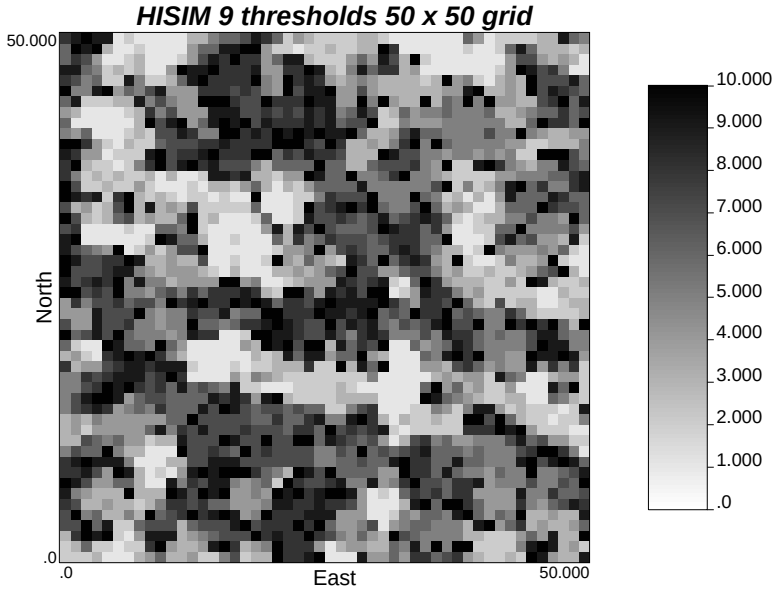

Figure 1: Map showing the result for the original implementation of HISIM. Higher thresholds present high nugget effect.

order until the sum of probabilities is higher than the random number multiplied by the total sum of probabilities. In this manner, nodes with higher probability of being below the threshold, i.e. with higher kriging estimates, will have a larger probability of being switched down or eroded.

The example shown on **Figures 1 and 2** show that the variogram models are not reproduced for higher thresholds, that is for the thresholds that were simulated first in the algorithm.

The initial idea of loosing freedom discussed in [3] from one threshold to the next was therefore a misinterpretation of the results because the indicator variograms were not labelled correctly. The proposed correction of using cokriging instead of kriging to calculate the probabilities did not work since the cokriging estimate is the same than the kriging estimate at the first threshold (in the unconditional case).

The nugget effect seen at the highest threshold is due to the small difference between the probability of a node uninformed ($p \in [0.9 - 1.0]$) and a node that has been informed, i.e. switched down ($p = 1.0$). This leads to a virtual random drawing of the nodes. This effect is less severe when a lower threshold is being simulated, since the difference between a node uncorrelated with the conditioning data and the others is larger, so the drawing in not random anymore. Although a very high nugget effect is still present, some correlation can be observed.

Proposed Approaches

Modifying the mean in SK

The first proposal is to modify the mean for kriging the indicators. A simple example with one threshold at the median is used to test this method. Although intuitively the mean used when simple kriging should be $F(z_1) = 0.5$, where z_1 is the only cutoff, several means were used. The

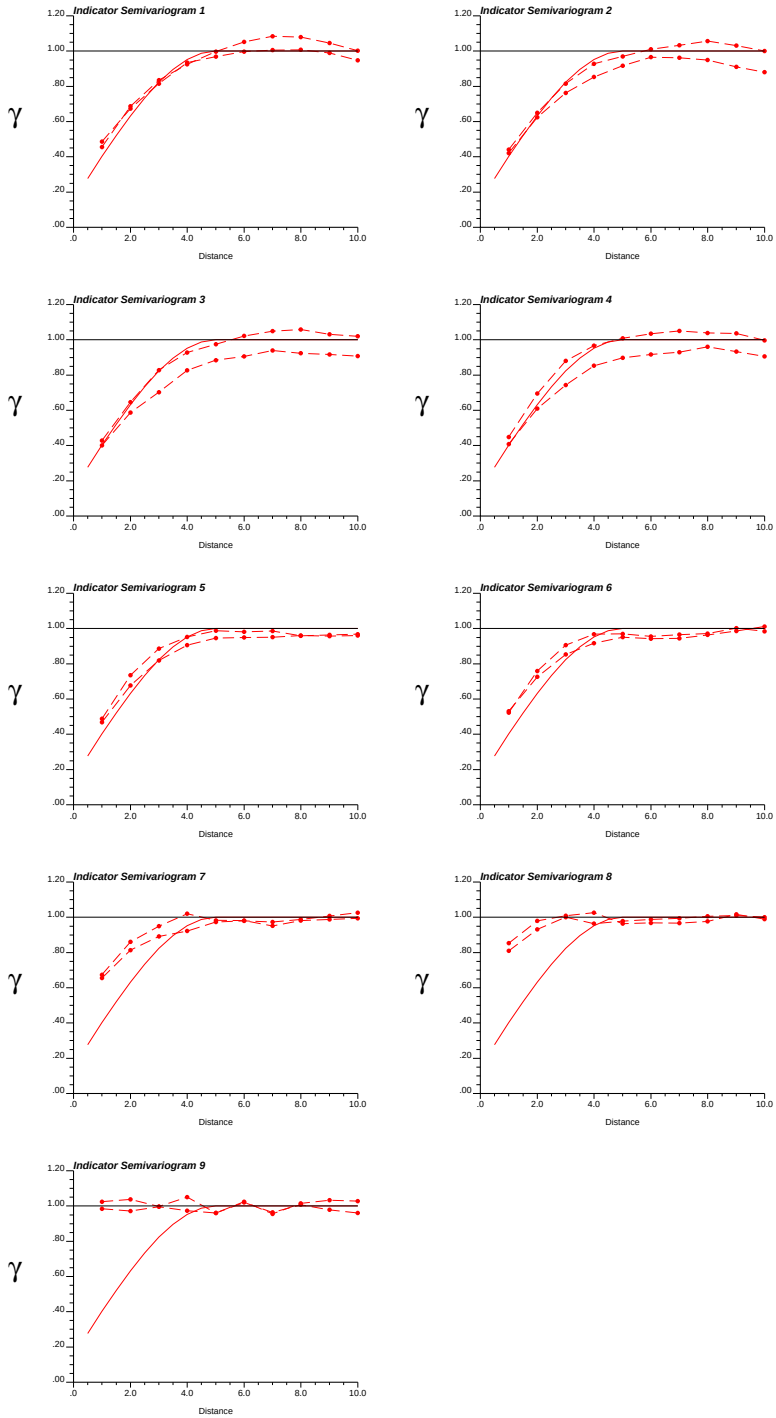

Figure 2: Variogram reproduction for the original implementation of HISIM.

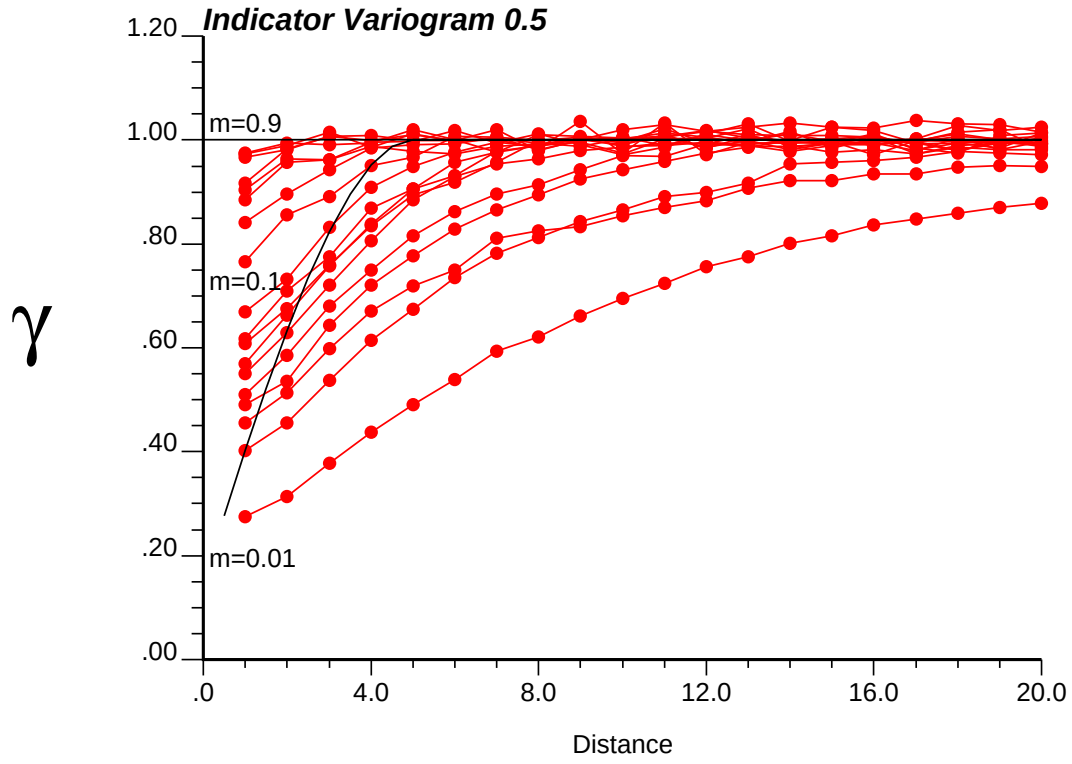

Figure 3: HISIM varying the simple kriging mean for a 1 threshold case (skmean varying from 0.01 to 0.9).

results are not encouraging, since, as seen in **Figure 3**, a decrease in nugget effect is accompanied by an increase in the correlation range. Therefore, the variogram cannot be reproduced by simply changing the simple kriging mean.

SIS hierarchical

The idea of eroding an initially high field is now replaced by the hierarchical application of SIS (sequential indicator simulation). The idea is to perform SIS at the highest threshold and then use the nodes simulated to be above that threshold as conditioning data for the following thresholds, since it is known that if the node is above a threshold, it is also above all other lower thresholds. This results in realizations that do not honor the proportions required, because of the bias introduced by the conditioning data. They are heavily biased towards zero, since those are the only nodes that can be used as conditioning data when proceeding from the highest threshold down. However, variogram reproduction was reasonable, except for the sill that depends on the proportion of ones and zeros (**Figure 4**).

This naturally leads to two ideas:

- To use an approach similar to the nested indicators proposed by Dagbert for kriging reserves [1].

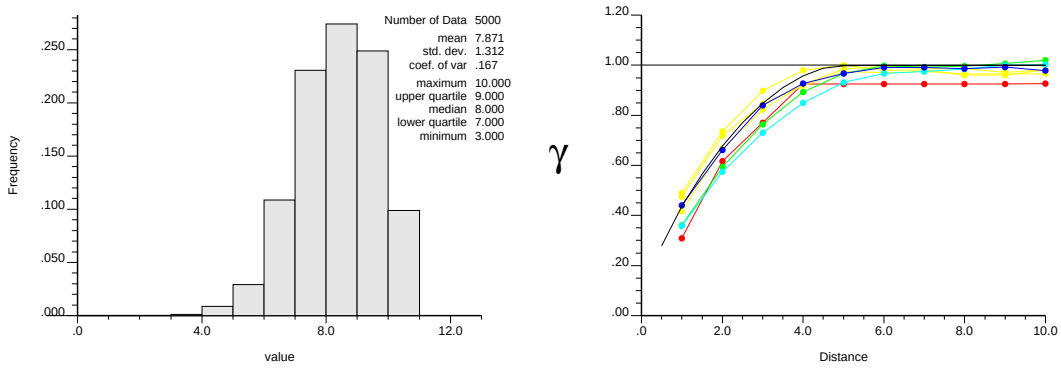

Figure 4: SISIM applied hierarchically. The use of zeros from the higher thresholds biases the conditioning data, generating realizations that do not honor the proportions. The histogram shows that there are no nodes being assigned to the lower thresholds, since they have all already been assigned to higher ones. The standardized variograms using the resulting proportions show that the correlation is preserved.

- To modify the proportions used as input to obtain the desired ones in the output.

Nested indicator simulation

The first solution was implemented with relative success. The steps involved in its implementation are:

1. At the highest threshold, the domain corresponds to all uninformed nodes.
2. An uninformed node is picked in the domain randomly.
3. The simple indicator kriging estimate at the current threshold is calculated given the nearby data and previously simulated nodes.
4. A random number is drawn and a one is assigned to the node if this random number is lower or equal than the simple indicator kriging estimate of the probability at that threshold, and a zero otherwise.
5. Go back to 2 until all nodes in the domain have been visited.
6. If the value is above the threshold, that is a value of zero was assigned in the binary simulation, then eliminate the node of the domain for the next threshold.
7. If the value is below the threshold include it in the domain for the next threshold.
8. Repeat for all thresholds.

In the end, a continuous value can be assigned at every node, since the class to which it belongs is known. The usual interpolation and extrapolation beyond the discrete cumulative distribution function used in SIS is required (see for example [2]).

One of the problems of this approach is that correlation between thresholds is not imposed, therefore the result looks patchy, and it is common to find high values beside low values without the appropriate transition in between. This algorithm has been fully developed. Refer to [4] for further details and applications.

Correcting the proportions: Markov and empirical approaches

The second proposed solution attempts to account for the bias generated by the conditioning data. The question is: How much do we have to change the input proportion to obtain the required proportions?

After several attempts, a correction factor for the proportion used as a mean was applied. This implies a non-linear additive correction to the estimated probabilities. Consider the original estimate, using P_{Theo}, and the new estimate using P_{Corr}.

$$\begin{array}{rcl} [i(\mathbf{u})]^*_{Theo} & = & \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u}) \cdot i(\mathbf{u}_\alpha) + [1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u})] \cdot P_{Theo} \\ [i(\mathbf{u})]^*_{Corr} & = & \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u}) \cdot i(\mathbf{u}_\alpha) + [1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u})] \cdot P_{Corr} \end{array}$$

The difference in the estimate is:

$$\Delta = [1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}(\mathbf{u})] \cdot (P_{Corr} - P_{Theo})$$

Next to a data location, this factor vanishes, since, the sum of the kriging weights approaches one. On the other hand, far from data, this factor tends to its maximum, $P_{Corr} - P_{Theo}$.

Notice also that the same type of correction would be possible using a cokriging approach. The correlation between indicators at different thresholds does not need to be input. It can be calculated, given the proportions of ones at the current threshold p_2, and the proportion of ones at the previous (higher) threshold p_1:

$$\rho = \sqrt{\frac{p_2 \cdot (1 - p_1)}{p_1 \cdot (1 - p_2)}}$$

We experimented also with this approach. Results showed that the proportions were not reproduced either. Variograms showed a small increase on the nugget effect for lower thresholds, i.e. the last ones being simulated. However, the range of correlation was preserved (**Figure 5**).

An empirical correction factor for the simple kriging mean was found that "updates" the mean for every threshold. It is the ratio of the average probability expected for each node at the current threshold, over the average probability calculated considering the conditioning data:

$$f = \frac{p_2}{\frac{\sum_{i=1}^{nx} i^*_{SK}}{nx}}$$

where p_2 is the proportion at the current threshold, nx is the total number of nodes, and i^*_{SK} are the simple kriging estimates of the probabilities of being below the threshold. The simple kriging mean is then multiplied by this factor every time a new threshold is being simulated.

The results show a good reproduction of the histogram, but an increase in correlation for lower thresholds, along with a decrease in nugget effect as the simulation proceeds (**Figure 6**).

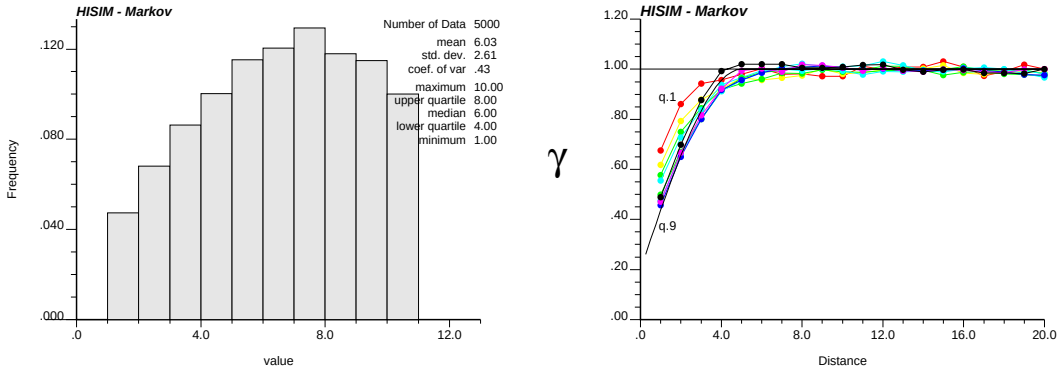

Figure 5: Hierarchical application of SIS using a Markov assumption for collocated cokriging of the indicators using the value at the previous higher threshold. The histogram is not reproduced (uniform distribution) and some increase in the nugget effect can be seen for lower thresholds.

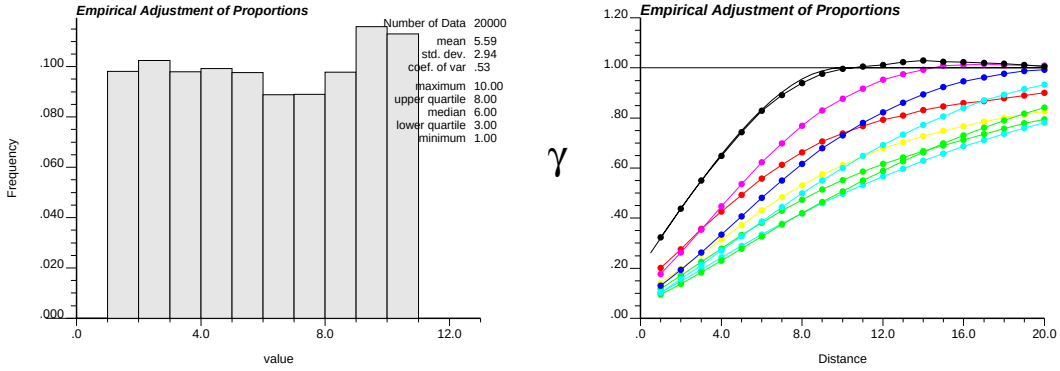

Figure 6: Empirical adjustment of the proportion to apply SIS hierarchically. Histogram reproduction is good, variograms show an increase in correlation and reduction in nugget effect.

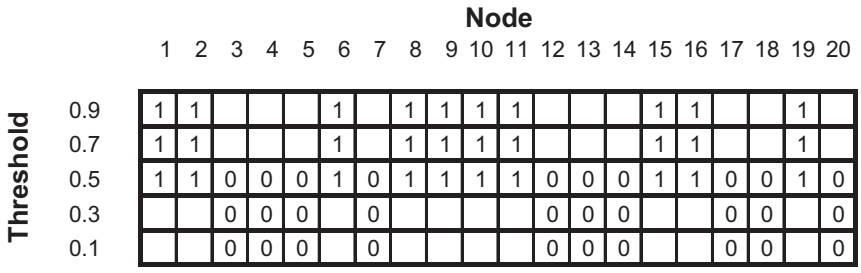

Figure 7: Illustration of median hierarchical indicator simulation. The nodes with a value higher than the median are used as conditioning data for lower thresholds, and the nodes with values below the median are used as conditioning data for all thresholds higher than the median.

Median Hierarchical Indicator Simulation

One last idea proposed is the use of SIS to simulate at the median, and then proceed up and down using the original hierarchical idea, that is, eroding in both directions, keeping the nodes set below the median when going to higher thresholds, or the nodes set above the median when going to lower thresholds. This is illustrated in **Figure 7**.

This algorithm proceeds as follows:

- Simulate by SIS (or any other binary simulation method, such as truncated Gaussian simulation) the median threshold. Every node is assigned a one or a zero, depending if they are below or above the median value, respectively.
- For thresholds below the median:
  - Use the nodes set above the median, that is those coded with a zero, as conditioning data.
  - Calculate the simple indicator kriging estimates at every location.
  - Select one location by Monte Carlo drawing, using the probabilities previously calculated by simple indicator kriging.
  - Repeat until the right proportion of nodes has been set below the current threshold.
  - Set all the nodes that have not been switched as zero. Their values are between the current cutoff and the higher threshold.
  - Use all the nodes with zero values (the ones that have just been coded and those that were coded in a previous threshold simulation) as conditioning data for the next threshold.
  - Repeat until the lowest threshold has been simulated.
- For nodes above the median:
  - Code the data above the median as ones and the nodes with values below the median as zeros.
  - Proceed as with the thresholds below the median, but working with the probability of being above the cutoff, instead of below it.

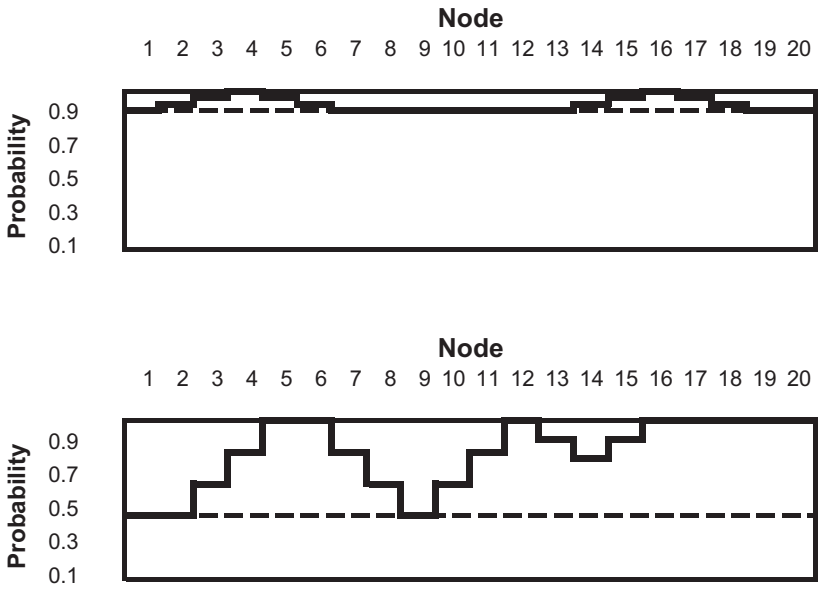

Figure 8: Illustration of the case when drawing nodes by Monte Carlo simulation is virtually random (top) and when the drawing is effective and accounts for the now larger differences in probabilities (bottom).

The algorithm is symmetric with respect to the median. Results showed good reproduction of the histogram: the number of nodes above and below the median presents ergodic fluctuations from SIS or the algorithm used to generate this binary simulation. The proportions for other thresholds is guaranteed by construction since the number of nodes to switch is defined by the proportions. Variogram reproduction at the median is also obtained depending on the algorithm used to generate the initial binary simulation. At other thresholds, variogram reproduction is obtained just as in the original case, but here, the problem of having a small difference between the probability calculated by simple kriging and the probability for nodes away from data is large, so the drawing of the nodes to be switched is not random (**Figure 8**).

A first example is shown in **Figure 9**. Twenty realizations of a one dimensional array of 3000 nodes with intrinsically correlated indicators, that is the so called mosaic random function model, was tested. Nine thresholds and a spherical variogram with a range of 10 units and a 10% of nugget effect was used. The results are encouraging. Variogram reproduction is good, although a slight increase in correlation can be seen for indicators far away from the median.

A second example with a multivariate Gaussian variable is also presented (**Figure 10**). In this case variogram reproduction is poor at thresholds other than the median. However, the range of correlation is preserved. Again histogram is reproduced by construction.

Finally, a non-Gaussian variable was used (**Figure 10**). The results are a mixture of the previous two examples. Good reproduction of the indicator variograms at some thresholds and poor reproduction at others.

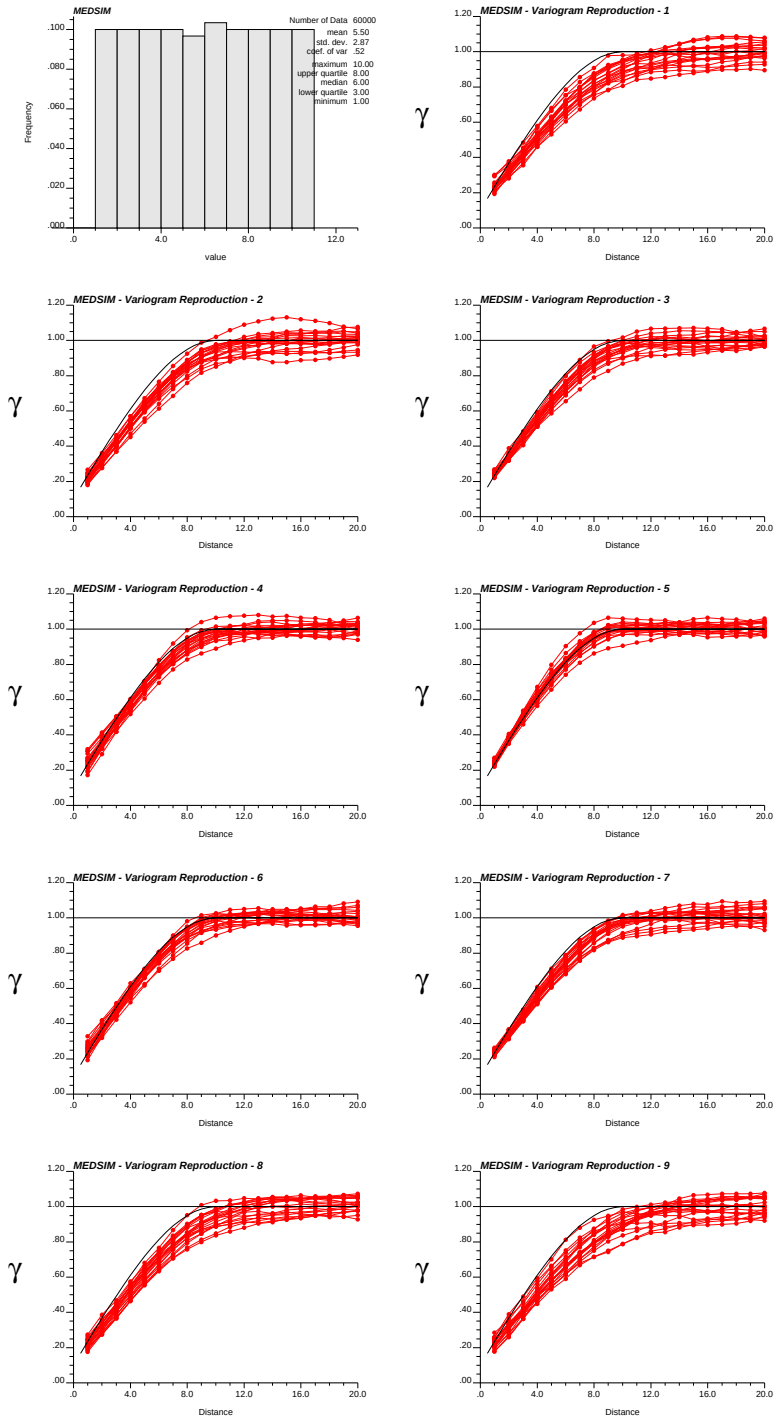

Figure 9: Application of Median Hierarchical Indicator simulation for an intrinsically correlated variable or mosaic model.

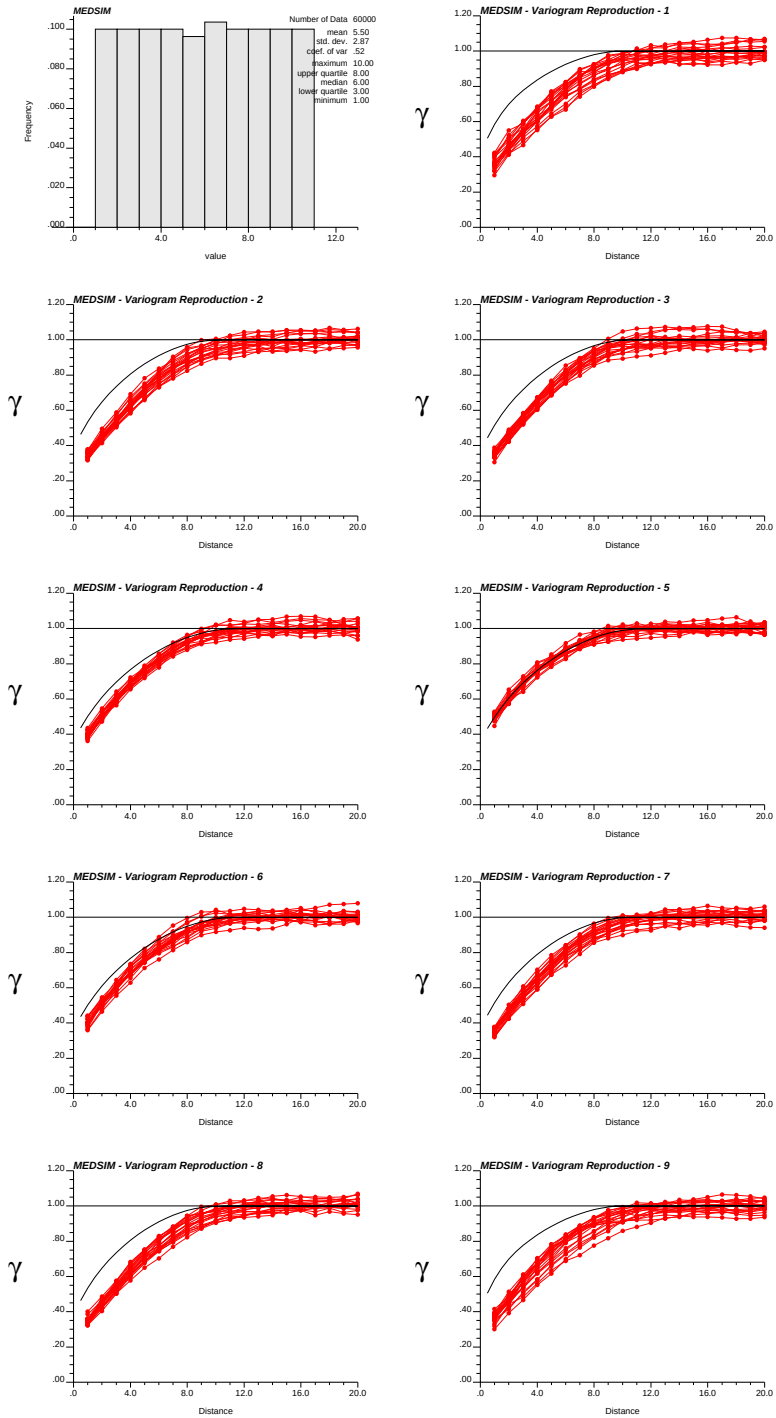

Figure 10: Application of Median Hierarchical Indicator simulation fopr a multi-Gaussian variable.

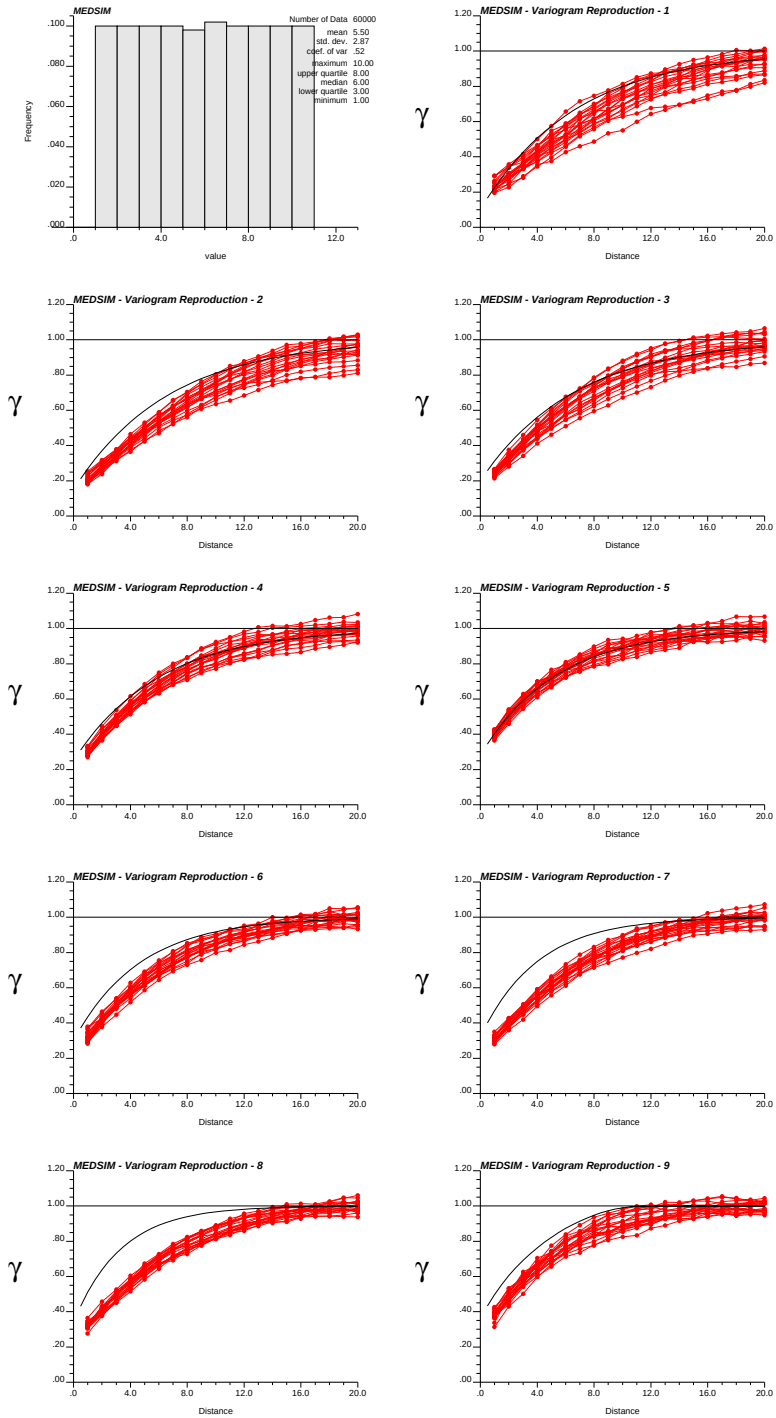

Figure 11: Application of Median Hierarchical Indicator simulation for a non-Gaussian variable.

Conclusions

Simulating one threshold at a time is appealing since this avoids order relation deviations and permits a useful framework for incorporating multiple-point statistics.

The original idea failed in that correlation could not be reproduced for high thresholds. Correlation was recovered as the algorithm proceeded to the lower thresholds. The use of another technique such as SIS for locking the realization at a given threshold was explored, however, variogram reproduction was never achieved in a completely satisfactory way. Apparently, the biased conditioning generates unavoidable bias in the covariance reproduction. This problem is difficult to tackle, since we proceed sequentially, and this generates a constant change in the magnitude of the bias. The idea of correcting while simulating could be a possible way to fix this problem.

Among all the techniques explored, the nested approach seems reasonable, because it rests in the well known indicator approach. Research could focus on correcting for the increase in nugget effect generated by not accounting for the zeros from the higher thresholds. The result would be different than the one obtained through SIS, since the nested approach would generate a map that truly resembles a mosaic, in the sense that patches of different classes would be randomly distributed in the field.

As a final comment, the incorporation of multiple-point statistics could be approached separately from this hierarchical algorithm. Runs could be drawn directly into a field without even considering the two-point statistics.

References

[1] M. Dagbert. Nested indicator approach for ore reserve estimation in highly variable mineralization. In *92nd Annual General Meeting of CIM - 1990*, Ottawa, Ontario, May 1990.

[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] J. Ortiz C. Research note: HISIM - Hierarchical Indicator Simulation. In *Centre For Computational Geostatistics*, volume 3, Edmonton, AB, 2001.

[4] J. Ortiz C. and C. V. Deutsch. Hierarchical indicator simulation. In *Proceedings of the 30th International APCOM Symposium*, Phoenix,AZ, February 2002. Society of Mining Engineers.