Estimating Conditional Moments with Inverse Distance Weights in Sequential Gaussian Simulation

Olena Babak and Clayton V. Deutsch

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

Application of Bayes Law under a multivariate Gaussian distribution leads to the normal equations or simple kriging as the correct approach to calculate conditional moments. These moments (the mean and variance) fully define the conditional distribution of uncertainty. The multivariate Gaussian distribution makes strong assumptions of stationarity and ergodicity and there are circumstances when alternate schemes can be considered to estimate the conditional moments. Inverse distance (ID) weights is a robust estimation approach. This note describes how ID weights can be applied in sequential Gaussian simulation. Of particular concern is avoiding the introduction of a bias. This is discussed with example.

Introduction

The sequential approach to simulation is grounded in the theory decomposing the joint multivariate distribution into a succession of conditional distributions by recursive application of Bayes Law. Conditional distributions are established by the normal equations or simple kriging. The problem is that real regionalized variables do not necessarily follow the neat theory of the Gaussian distribution: perfect Gaussian stationary variable in an infinite domain. We often modify our approach to estimate the conditional moments to impart desirable features in simulation. These modifications must be approached carefully because we may impart biases – especially in the back transformed histogram.

This short note presents the use of inverse distance (ID) for the calculation of conditional moments. ID is remarkably robust: it has no string effect, no negative weights and straightforward control on smoothness. It does not handle clustered data very well and does not minimize a clearly defined measure of error variance; however, there may be cases where the pros outweigh the cons.

Inverse Distance Interpolation

An inverse distance interpolation is one of the simplest and most popular interpolation techniques. It combines the proximity concept with the gradual change of the trend surface. An inverse distance (ID) weighted interpolation is defined as a spatially weighted average of the sample values within a search neighborhood. It is calculated as

$$Z^*(u) = \sum_{i=1}^{n} \lambda_i Z(u_i), \tag{1}$$

where **u** is the estimation location, $u_i,\ i=1,\ldots,n,$ are the locations of the sample points within the search neighborhood, $Z^*(u)$ is the inverse distance estimate at the estimatiom location, *n* is the number of sample points, $\lambda_i,\ i=1,\ldots,n,$ are the weights assigned to each sample point, and $Z(u_i),\ i=1,\ldots,n,$ are the conditioning data at sample points. The weights are determined as

$$\lambda_i = \frac{\left(\dfrac{1}{d_i^{\,p}}\right)}{\sum\limits_{i=1}^{n}\left(\dfrac{1}{d_i^{\,p}}\right)}, \quad (i = 1,\ldots,n), \tag{2}$$

where d_i are the Euclidian distances between estimation location and sample points, and exponent *p* is the power or distance exponent value. The most common value applied for the power p is 2; then estimator in

(1)-(2) is called inverse squared distance (ISD) interpolator. However, any value for *p* can be chosen by user. As *p* increases, the interpolated value by inverse distance is assigned the value of the nearest sample point, that is, inverse distance estimate becomes the same as estimate produced by polygonal method. (Diadato and Ceccarelli, 2005; Mueller et al., 2005). The advantage of the inverse distance technique is that it can be easily applied in any number of dimensions and provide reasonable estimates. Several modifications of the inverse distance include gradient inverse distance interpolation (GIDW), anisotropic inverse distance interpolation (AIDW), etc. (Price et al., 1998; Nalder and Wein, 2000)

Mean and Variance of the Inverse Distance Estimator

The mean and variance of the Inverse distance estimator $Z^*(u)$ at estimation location **u** given by (1)-(2) can be calculated under stationarity (that is, $E(Z(u)) = m, Var(Z(u)) = \sigma^2, \forall u)$) to be equal to:

$$\begin{aligned} E(Z^*(u)) &= E\left(\sum_{i=1}^{n}\lambda_i Z(u_i)\right) = \sum_{i=1}^{n}\lambda_i E(Z(u_i)) = m\sum_{i=1}^{n}\lambda_i \\ &= m\sum_{j=1}^{n}\frac{\left(\dfrac{1}{d_j^p}\right)}{\sum_{i=1}^{n}\left(\dfrac{1}{d_i^p}\right)} = m; \end{aligned} \tag{3}$$

$$\begin{aligned} Var(Z^*(u)) &= Var\left(\sum_{i=1}^{n}\lambda_i Z(u_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Cov(Z(u_i),Z(u_j)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\left(\dfrac{1}{d_i^p}\right)}{\sum_{k=1}^{n}\left(\dfrac{1}{d_k^p}\right)}\frac{\left(\dfrac{1}{d_j^p}\right)}{\sum_{l=1}^{n}\left(\dfrac{1}{d_l^p}\right)}\right]Cov(Z(u_i),Z(u_j)) \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\dfrac{1}{d_i^p}\right)\left(\dfrac{1}{d_j^p}\right)Cov(Z(u_i),Z(u_j))}{\left[\sum_{k=1}^{n}\left(\dfrac{1}{d_k^p}\right)\right]^2}; \end{aligned} \tag{4}$$

where d_i are the Euclidian distances between estimation location and sample points, exponent *p* is the power or distance exponent value used in inverse distance interpolation and $Cov(Z(\mathrm{u}_i), Z(\mathrm{u}_j)), \quad i, j = 1,\cdots,n,$ denotes data-to-data covariance function calculated under assumption of stationarity though the semivariogram model $2\gamma(h)$ $(Cov(Z(\mathrm{u}_i), Z(\mathrm{u}_j)) = C(u_i, u_j), \quad i, j = 1,\cdots,n)$. The estimate and variance of the Inverse distance estimator at the data location are set to the data value at that location and stationary domain variance σ^2, respectively.

The map of the inverse distance estimates is smooth; the smoothing is especially strong far from the data locations. The smoothing effect of inverse distance interpolation technique is directly related to the IDW variance via this expression

$$\text{Smoothing effect} = \sigma^2 - Var(Z^*(u)) = \sigma^2 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\dfrac{1}{d_i^p}\right)\left(\dfrac{1}{d_j^p}\right)Cov(Z(u_i),Z(u_j))}{\left[\sum_{k=1}^{n}\left(\dfrac{1}{d_k^p}\right)\right]^2} \tag{5}$$

Note that smoothing effect of the inverse distance interpolator in (5) is also referred to as the missing variance. This is because by adding the variable with variance (5) to the inverse distance estimate we will obtain new variable with variance equal to the stationary domain variance σ^2. Note that at the data locations missing variance is equal to zero.

Inverse Distance Non-sequential Simulation

The algorithm for the inverse distance simulation is the following:

- Transform the original variable into the normal space, $Z(u)\xrightarrow{N}Y(u)$.
- Go to a location u in the study domain and calculate inverse distance estimate and corresponding inverse distance estimation variance based on data from the search neighborhood:

$$Y^*(u)=\frac{\sum_{j=1}^{n}\left(\frac{1}{d_j^p}\right)Y(u_j)}{\sum_{i=1}^{n}\left(\frac{1}{d_i^p}\right)}, \tag{6}$$

$$Var(Y^*(u))=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{d_i^p}\right)\left(\frac{1}{d_j^p}\right)C(u_i,u_j)}{\left[\sum_{k=1}^{n}\left(\frac{1}{d_k^p}\right)\right]^2}. \tag{7}$$

- Draw a random residual $R(u)$ from a normal distribution with zero mean and variance from (5).
- Obtain the inverse distance simulated value as $Y_s(u) = Y^*(u) + R(u).$ Note that the simulated values are characterized by the stationary mean and variance

$$E(Y_s(u)) = E(Y^*(u) + R(u)) = E(Y^*(u)) + E(R(u)) = 0, \tag{8}$$

$$Var(Y_s(u)) = Var(Y^*(u) + R(u)) = Var(Y^*(u)) + Var(R(u))$$

$$=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{d_i^p}\right)\left(\frac{1}{d_j^p}\right)Cov(Z(u_i),Z(u_j))}{\left[\sum_{k=1}^{n}\left(\frac{1}{d_k^p}\right)\right]^2}+\sigma^2-\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{d_i^p}\right)\left(\frac{1}{d_j^p}\right)Cov(Z(u_i),Z(u_j))}{\left[\sum_{k=1}^{n}\left(\frac{1}{d_k^p}\right)\right]^2}=\sigma^2.$$

- Visit all locations and repeat the procedure above.
- Back-transform all data values and simulated values.
- Create any desired number of realizations by repeating with different number seed. A different seed results in different residuals for each simulated node.

Pre-and post-processing (transforming to and back transforming from normal space) is required in the inverse distance simulation for the same reason it is required in sequential Gaussian simulation. That is, the Gaussian distribution is used to insure that the distribution of simulated values is correct.

Inverse Distance Non-sequential Simulation: Example

To illustrate the performance of the inverse distance non-sequential simulation the following small example was considered. A well known GSLIB data set 'cluster.dat' is selected for analysis. That data consists of about 100 data that are sampled on a random stratified grid and 40 data that are clustered in high valued areas. We discard the clustered data. The 2-D area of interest is 50 by 50 distance units. The distribution of data is approximately lognormal with a mean of 2.5 and a standard deviation of 5.0. The spatial continuity of the data is described by isotropic Spherical variogram model with range of correlation 10 and nugget effect of 0.3. Figure 1 shows the location of the 100 data as well as one example inverse distance non-sequential simulation realization. When simulating maximum 20 closest original conditional data were used. Clearly realization does not preserve much continuity structure inherent in the data, despite realizations met closely both target mean and variance (0 and 1, respectively) requirements, see Figure 1. Figure 2 shows for comparison example realization of the non-sequential Gaussian simulation. Result is pretty much the same as was obtain via non-sequential Inverse Distance simulation, that is, despite realization meet target statistical requirement; there is not much continuity structure in the generated realization. Figure 2 also shows one sequential Gaussian realization (maximum 20 closest simulated nodes and data combined were used in simulation), clearly the realization shows clear spatial structure; moreover, it also meets target statistical requirements.

Inverse Distance Sequential Simulation

It was shown that inverse distance simulation needs to be performed in sequential mode. That is, not only data, but also simulated nodes need to be used when building conditional distribution with inverse distance method. Figure 3 shows example realization obtained using inverse distance sequential simulation with 5, 10 and 20 closest simulated nodes and original conditioning data combined. Figure 3 also shows for comparison non-sequential inverse distance realization. It is apparent from Figure 3 that sequential simulation with inverse distance nicely preserves the continuity structure much better. Moreover, it can be also noted that with increase in the number of data (simulated and original), simulation becomes more pixelated. Figure 4 show the reproduction of the target mean of zero and target variance of 1 in the sequential inverse distance simulation. It can be seen from Figure 4 that sequential inverse distance simulation with 5 data meets the target statistics almost exactly, while sequential inverse distance simulation with 10 and 20 data results in slight variance inflation and variance under-inflation. The reason for the variance mismatch is simple. As in the case of ordinary kriging, collocated cokriging, the inverse distance sequential simulation has no control over the reproduction of the covariance structure. Specifically, the covariance between two simulated values is not reproduced since

$$Cov(Y_s(u_1),Y_s(u_2)) = Cov(Y^*(u_1)+R(u_1),Y^*(u_2)+R(u_2)) = Cov(Y^*(u_1),Y^*(u_2))$$

$$= Cov\left[\frac{\sum_{j=1}^{n}\left(\frac{1}{d_{1j}^{p}}\right)Y(u_j)}{\sum_{i=1}^{n}\left(\frac{1}{d_{1i}^{p}}\right)}, \frac{\sum_{l=1}^{n}\left(\frac{1}{d_{2l}^{p}}\right)Y(u_l)}{\sum_{k=1}^{n}\left(\frac{1}{d_{2k}^{p}}\right)}\right] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{d_{1i}^{p}}\right)\left(\frac{1}{d_{2j}^{p}}\right)Cov(Y(u_i),Y(u_j))}{\sum_{k=1}^{n}\left(\frac{1}{d_{1k}^{p}}\right)\sum_{l=1}^{n}\left(\frac{1}{d_{2l}^{p}}\right)}$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{d_{1i}^{p}}\right)\left(\frac{1}{d_{2j}^{p}}\right)Cov(u_i,u_j)}{\sum_{k=1}^{n}\left(\frac{1}{d_{1k}^{p}}\right)\sum_{l=1}^{n}\left(\frac{1}{d_{2l}^{p}}\right)} \neq C(u_1,u_2).$$

Relying on the assumption that $Cov(Y_s(u_1),Y_s(u_2)) = C(u_1,u_2),$ and building new conditional distributions using not only data, but also simulated nodes can result in estimates which are over- or under-dispersed. Therefore, when applying sequential inverse distance simulation special care must be taken in choosing simulation parameters.

Figure 5 shows the reproduction of the target variogram model by the non- sequential inverse distance simulation and sequential inverse distance simulation with 5, 10 and 20 simulated nodes and original conditional data combined. No multiple grid search was used. In general, we can observe that sequential inverse distance simulation ensures reproduction of the target variogram at small lag distances better when 10 simulated nodes and original conditional data combined are used in simulation.

Figure 6 shows influence of the number of original conditioning data on the sequential inverse distance simulated realization. Several cases for the number of original conditioning data are considered, 5, 10, 15 and 20, maximum number of simulated nodes is set to 5; maximum search radii were set to the size of the domain. We can observe from Figure 6 that with increase in the number of original data used in simulation apart from the number of simulated nodes, there appear to be more disconnectedness and pixilation in the generated realization. As result, we can conclude that only local data should be preferred for sequential inverse distance simulation.

Conclusions

In this paper a new modification to the sequential Gaussian simulation was proposed. The modification was based on the use of inverse distance (ID) to calculate the conditional moments. The applicability of the proposed method, referred to as inverse distance simulation, was illustrated via small example.

References

Deutsch, CV and Journel, A.G.: GSLIB: Geostatistical Software Library and Users Guide,Oxford University Press, New York, second edition, 1998.

Diadato, N. and Ceccarelli: Interpolation Processes using Multivariate Geostatistics for Mapping of Climatological Precipitation Mean in the Sannio Mountains (southern Italy), Earth Surface Processes and Landforms, 2005.

Mueller, T.G., Dhanikonda, S.R.K., Pusuluri, N.B., Karathanasis, A.D., Mathias, K.K., Mijatovic, B. and Sears, B.G.: Optimizing Inverse Distance Weighted Interpolation with Cross-Validation, Soil Science, 2005.

Nalder, I.A., Wein, R.W. Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest, Agric. For. Meteorol, 1998.

Price, D.T., McKenney D.W., Nalder I.A., Hutchinson M.F. and Kesteven J.L: A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data, Agricultural and Forest Meteorology, 2000.

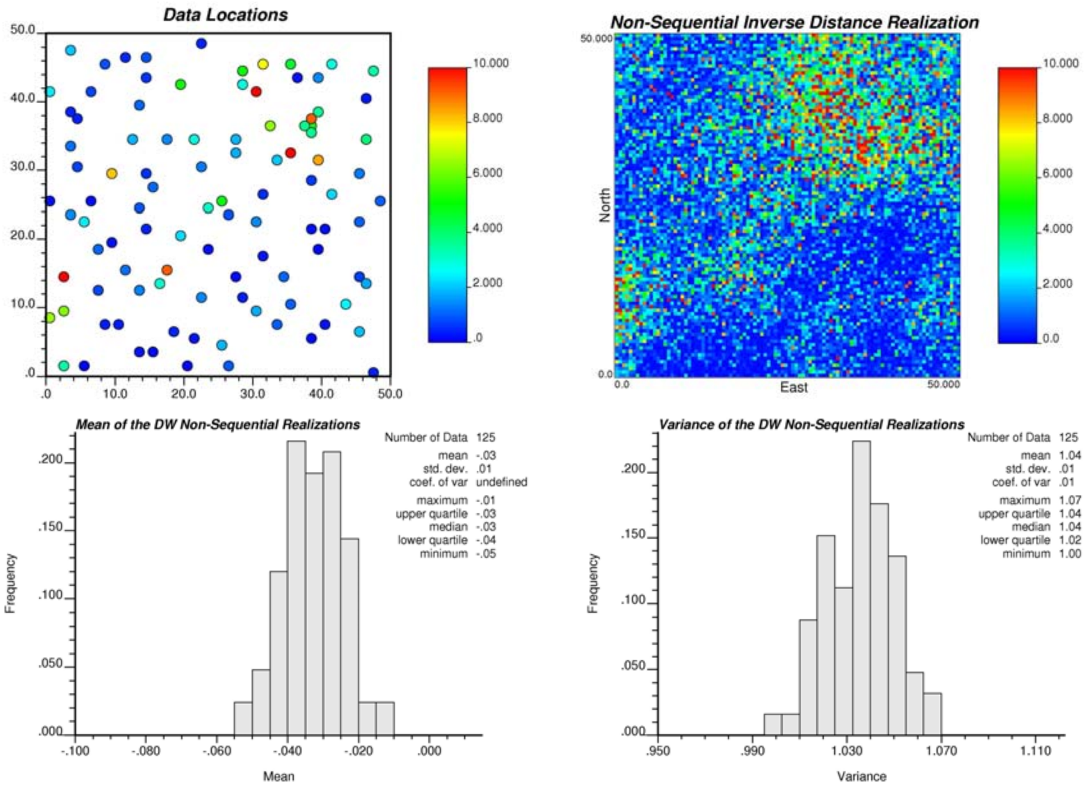

Figure 1: Locations of the 100 data (top left); Example inverse distance non-sequential simulation realization (top right); Means (bottom left) and variances (bottom eight) of 100 realizations generated via non-sequential inverse distance approach.

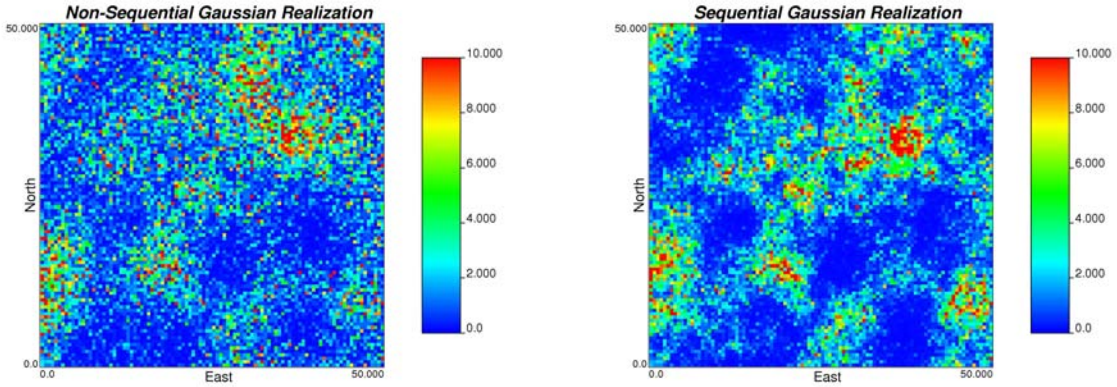

Figure 2: Example realization of the non-sequential Gaussian simulation (left) and example sequential Gaussian realization (right).

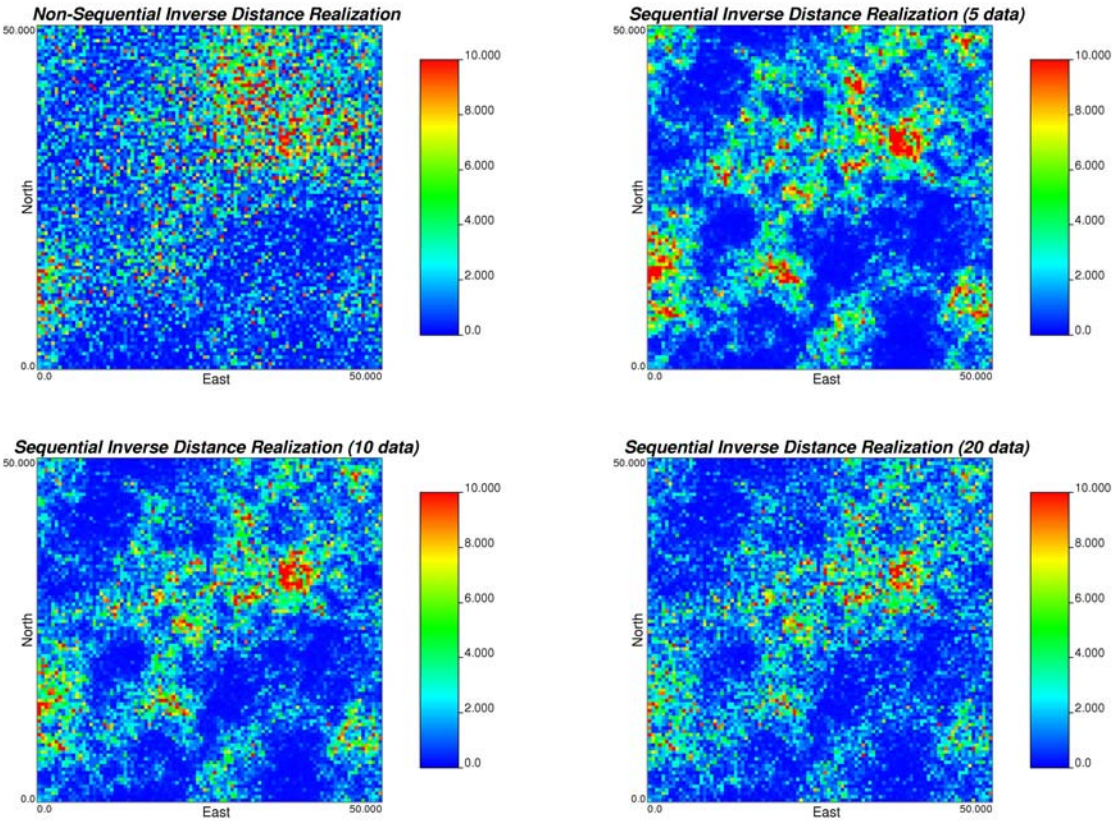

Figure 3: Example inverse distance sequential realization obtained with only original conditioning data (top left), with maximum 5 conditioning data and simulated nodes combined (top right), with maximum 10 conditioning data and simulated nodes combined (bottom left) and with maximum 520conditioning data and simulated nodes combined (bottom right).

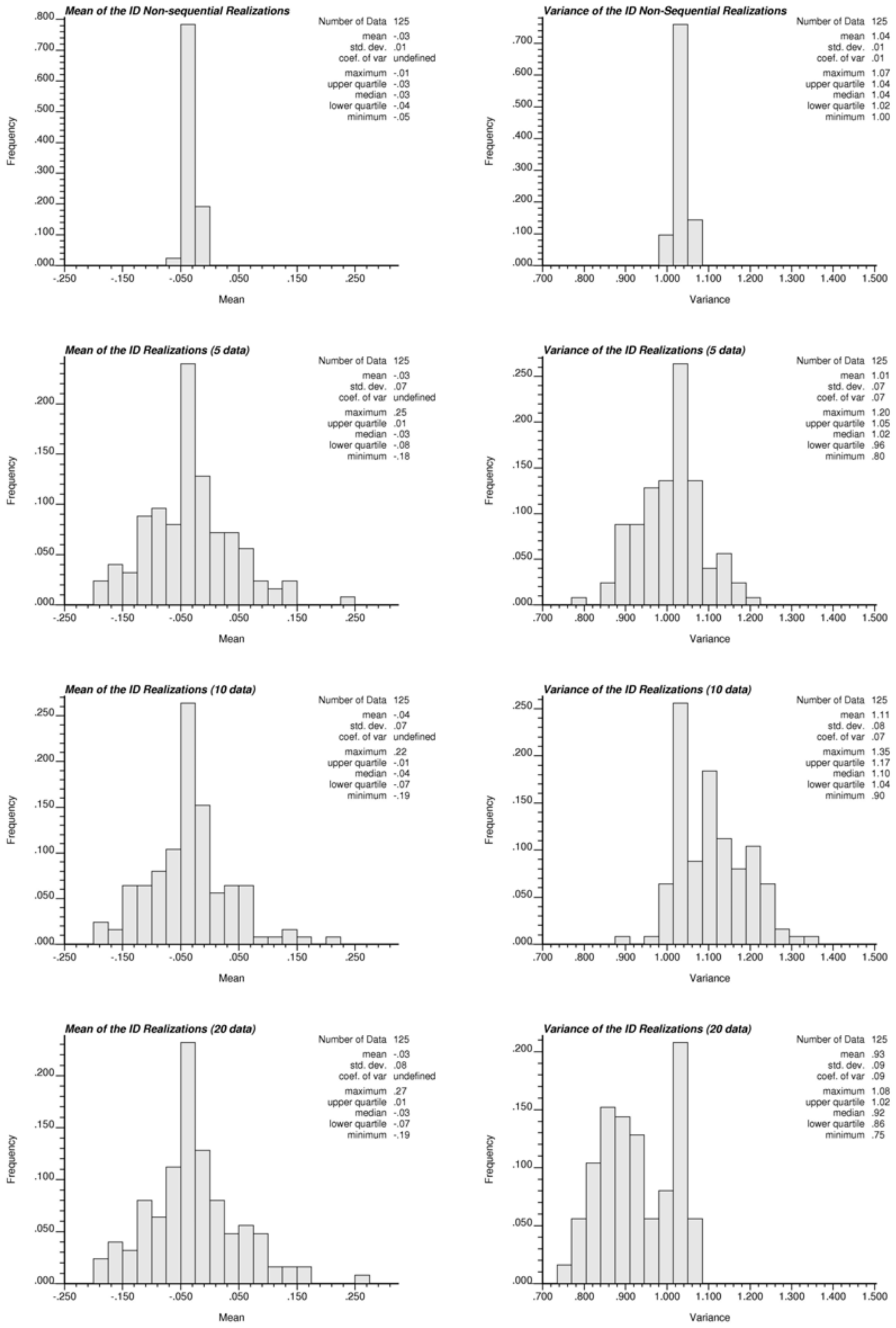

Figure 4: Reproduction of the target mean of 0 (left) and target variance of 1 (right) by the non-sequential inverse distance simulation and sequential inverse distance simulation with maximum 5(top), 10 (middle) and 20 (bottom) closest simulated nodes and data combined.

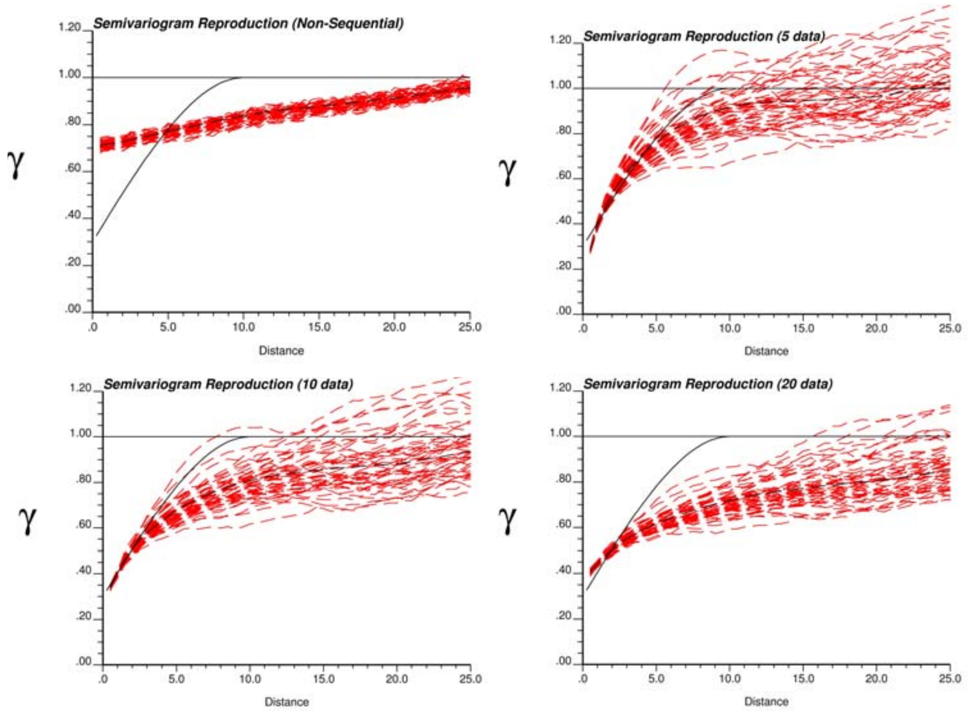

Figure 5: Reproduction of the target variogram by the non-sequential inverse distance simulation (top left) and the sequential inverse distance simulation with 5(top right), 10 (bottom left) and 20 (bottom right) simulated nodes and original conditioning data combined. Rep dashed lines are experimental variograms calculated from the simulated realizations; black dashed lines are average experimental variograms calculated from the simulated realizations; and black solid line is variogram model fitted to the original 100 data.

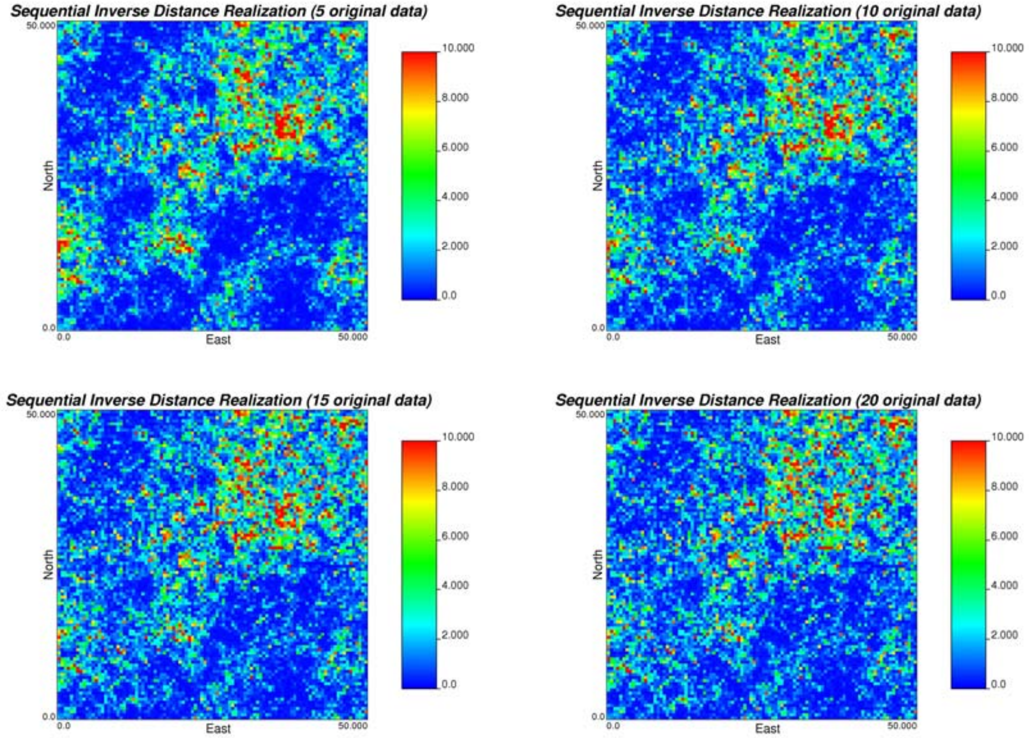
Sequential Inverse Distance Realization (5 original data)
Sequential Inverse Distance Realization (10 original data)
Sequential Inverse Distance Realization (15 original data)
Sequential Inverse Distance Realization (20 original data)
North
East
0.0
50.000
10.000
8.000
6.000
4.000
2.000
0.0

Figure 6: Influence of the number of original conditioning data on the sequential inverse distance simulated realization: 5 data (top left), 10 data (top right), 15 (bottom left), 20 data (bottom right).