DSSIM-HR: A FORTRAN 90 Program for Direct Sequential Simulation with Histogram Reproduction

Bora Oz (boz@gpu.srv.ualberta.ca)
University of Alberta, Edmonton, Alberta

Clayton V. Deutsch (cdeutsch@civil.ualberta.ca)
University of Alberta, Edmonton, Alberta

Thomas. T. Tran (tttr@chevron.com)
Chevron Petroleum Technology Company, San Ramon, California

YuLong Xie (YuLong.Xie@pnl.gov)
Pacific Northwest National Laboratory, Richland, Washington

Abstract

Sequential simulation is a commonly used geostatistical simulation technique. The most widely used version of this technique is sequential Gaussian simulation (SGS), where the data are transformed to follow a Gaussian distribution and the entire multivariate distribution is then assumed to be Gaussian. This critical assumption greatly simplifies the simulation process. Data of different volumetric support cannot be used because most variables do not average linearly after Gaussian transformation. The idea of direct sequential simulation (DSS) is to avoid the Gaussian transform and permit integraton of multiscale data. A longstanding problem of direct simulation is that the histogram of the variable is not reproduced even though the mean, variance, and variogram are reproduced. This is due to the unknown shape of the conditional distributions. We derive a simple and theoretically valid approach to establish the conditional distribution shapes to ensure histogram reproduction and valid distributions of uncertainty. The new approach has been coded in a FORTRAN 90 program called `DSSIM-HR`*, where the extension* `HR` *corresponds to the feature of "Histogram Reproduction".*

Keywords: Geostatistical Simulation, Realizations, Multiscale Data, Gaussian Transformation

Introduction

Sequential simulation is a commonly applied geostatistical algorithm (Goméz-Hernández and Journel 1993; Isaaks 1990; Johnson 1987). Sequential simulation can be seen as Monte Carlo simulation from a multivariate distribution by decomposing that multivariate distribution into a succession of conditional distributions by recursive application of Bayes law. The multivariate Gaussian distribution is systematically applied to continuous variables because of its extremely favorable mathematical properties; the shape of *all* conditional distributions

are Gaussian with mean and variance given by simple (co) kriging. Real Z-data are never Gaussian; nevertheless, they can be transformed to a Y-Gaussian variable. Simulation is done in Y Gaussian space, and the simulated y values are back transformed to original z data units. Secondary data such as seismic data can also be used after transformation to a Gaussian distribution and assuming that both variables are jointly multivariate Gaussian. The sequential Gaussian simulation program (`SGSIM`) (Deutsch and Journel 1992) is one implementation of the algorithm.

There are some significant limitations to Gaussian transformation. For a given covariance, the Gaussian random function (RF) has maximum "disconnectedness" of extreme values; a property known as maximum entropy. Multivariate Gaussianity also entails that the pattern of spatial correlation is symmetric with respect to the median, that is, there is symmetric destructuration of extreme values.

Transformation of the data variable to a Gaussian distribution is problematic when dealing with data of different scale. Most variables average linearly (volumetric or mass proportions such as porosity) or with very particular known scaling laws (permeability). The non-linear transformation to a Gaussian variable means that the correct averaging in "Gaussian space" is complex and intractable. The Gaussian transformation must be avoided to permit multiscale data to be considered.

The motivation for a "direct" method is based on the requirement to simultaneously account for data of different volumetric scales. The notion of direct sequential simulaton was developed at the same time as sequential Gaussian simulation (Journel 1986). It was shown early in the development of sequential techniques that the variogram (covariance) structure and the global mean can be reproduced without transformation to Gaussian space provided that the simulated values are drawn from local conditional distributions centered at the simple (co)kriging estimates with a variance corresponding to the simple (co)kriging estimation variance. The conditional distribution could be of any shape. Exercising this freedom, however, leads to simulated realizations where the univariate histogram is not controlled and therefore not reproduced.

Until now, there has been no good way to decide what shape of distribution to use in DSS. Caers (2000a) used five different types of distributions and showed that there is no generalization on which type and shape of distribution is appropriate for a particular case; however, he showed that spike-type distributions with lower entropy characteristics might produce good results in some situations. Deutsch et. al. (2001) demonstrated that there is no single distribution-type that can be used throughout the simulation to reproduce the overall histogram.

The overall histogram is important; it is a first order statistic that has a first order affect on the calculations made with the simulated realizations. In fact, the main purpose of sequential simulation is reproduction of large-scale spatial characteristics such as the histogram and variogram structure. The inability of DSS to reproduce the input histogram has been a significant problem.

The same quantile-transformation procedure used to transform original values to a Gaussian distribution can be used to transform the output simulated values from DSS to the correct input histogram. The problem with this back transformation is that the final global histogram has no uncertainity (ergodic fluctuations) and, more importantly, large-scale data are no longer reproduced. The transformation can be modified so that local hard data are

reproduced; however, the problems of block data reproduction and statistical fluctuations in the input parameters remain.

Caers (2000a) proposed the use of "post processing" in order to transform the resulting simulated values into another set of variables which approximately reproduce the input global histogram. This post processing can destroy variogram reproduction and remove ergodic fluctuations. Caers (2000b) also proposed to reproduce the global histogram by formulating an objective function as a measure of difference between the input global histogram and the histogram of the simulated values. This objective function can be used to selectively accept/reject certain simulated values to ensure that the final realization reproduce the input global histogram. This approach can introduce artifacts and also removes ergodic fluctuations that are important for uncertainity analysis.

Soares (2000) proposed a different approach to reproduce the global histogram in DSS. The main idea of Soares's proposal is to sample from the global distribution using local simple (co)kriging estimates of the mean and variance. The simulated values are drawn from intervals of the global distribution, which are calculated with the local estimates of the mean and variance. The Gaussian transformation is used to determine the sampling intervals. Deutsch et. al. (2001) showed that sampling from the global distribution is not a good idea since the shape of the local distributions change considerably throughout the simulation.

Recently, Deutsch et. al. (2001) proposed an effective way for reproducing the input global histogram in DSS. The key ideas behind their method is to (1) work in original space, which is the whole motivation for DSS, and (2) work out the shape of the conditional distributions as function of their mean/variance using normal-score or Gaussian transformation. In each step of the direct sequential simulation, the idea is to identify the shape of the local distribution from a prepared database for different local (co)kriging mean and variance values. No data transformation is required and the overall histogram is guaranteed to be reproduced within statistical fluctuations.

The essence of this paper is to implement the proposal of (Deutsch et. al. 2001). The new program is coded in FORTRAN 90. The code and the operation of the program is in the style of GSLIB (Deutsch and Journel 1992). The approach and the implementation details are explained with illustrative examples.

Methodology

Let us consider a continuous variable Z with a known stationary global cdf $F_Z(z) = Prob\{Z < z\}$ and stationary variogram $\gamma_Z(\mathbf{h})$ at the original non-standard data scale. DSS of a continuous variable follows the classical steps of sequential simulation:

1. Randomly choose the spatial location of a node $\mathbf{u}$ from the grid nodes to be simulated.
2. Calculate the kriging estimate and estimation variance. These estimates at $\mathbf{u}$ are conditioned to the original data and all previously simulated values.
3. Draw a value from a distribution with the estimated mean and variance.
4. Return to step 1 until all nodes have been simulated.

The multiGaussian approach is typically used whereby the Gaussian or normal transformation is applied to the data before simulation and a back transformation is applied to the simulated values after simulation.

If the local distributions are centered at the simple (co)kriging mean and variance, any shape of distribution can be chosen and the stationary variogram model will be reproduced (Journel 1986). The ultimate histogram, however, will typically depend on the type and shape of the local distributions, the distribution of local data, and the normal distribution inherent to the central limit theorem. The central limit theorem is involved because of the averaging of kriging. Arbitrary choice of the local distribution shapes does not lead to reproduction of the global distribution.

The *correct* shape of the local distributions is known for the Gaussian case because we have a model for the full multivariate distribution. The original Z variable with stationary histogram $F_Z(z)$ could be transformed to a Gaussian Y variable with stationary standard normal distribution $G(y)$. The quantile or normal-score transformation is widely used for such transformation.

$$y = G^{-1}(F_Z(z)) \tag{1}$$

This transformation can be reversed at any time to get back to the original variable units:

$$z = F_Z^{-1}(G(y)) \tag{2}$$

The cumulative distribution functions $F_Z(z)$ and $G(y)$ and their inverse relations or quantile functions $F_Z^{-1}(z)$ and $G^{-1}(y)$ are known. Thus, we have direct link between Z and Y units. This transformation is unique, reversible, and non-linear.

Distributions of uncertainty in Z space can be determined from back transformation of non-standard univariate Gaussian distributions. The back transformation of the non-standard p^l-quantile:

$$z^l = F_Z^{-1}[G(G^{-1}(p^l)\sigma_k + y^*)] \tag{3}$$

where y^* and σ_k are the mean and standard deviation of the non-standard Gaussian distribution of uncertainty, and the $p^l, l = 1, .., L$ values are uniformly distributed between 0 and 1. The distribution of uncertainty in Z space is assembled from the $z^l, l = 1, ..., L$ values. There is no analytical expression for this distribution, aside from expression 3; nevertheless, the distribution is completely defined:

$$F_{Z,y^*,\sigma_k}(z) \tag{4}$$

The Gaussian parameters are added as subscripts in Equation 4 to denote a conditional distribution relating to a particular conditional distribution in Gaussian space. It is important to note that the shape of the z-conditional distributions are neither Gaussian nor identical to the original Z data distribution.

The shape of every z-conditional distribution is explicitly known (Equation 4) and the `DSSIM-HR` program uses these shapes to reproduce the global input histogram within statistical fluctuations.

The predetermined distribution shapes are used in the sequential simulation algorithm. All kriging and simulation is performed in original Z variable units. The Gaussian transform is only used to get the shape of the conditional distributions. In concept, the `DSSIM-HR` algorithm is very similar to the conventional sequential Gaussian simulation with the following modifications:

1. The mean and variance of the local distribution, $z^*(u)$ and $\sigma_z^2(u)$, are calculated in original Z units by simple (co) kriging using all relevant original data and previously simulated grid nodes or blocks.

2. The z-conditional distribution with the right z mean and variance (from step 1) is found by back transforming a non-standard Gaussian distribution.

3. A Z simulated value is drawn from this conditional distribution by Monte Carlo Simulation.

This approach will create realizations that reproduce the (1) local point and block data in the original Z data units, (2) the mean, variance, and variogram of the Z variable, and (3) the histogram of the Z variable. More details of this approach including a theoretical justification on how the input global histogram is reproduced is given by Deutsch et. al. (2001) and Tran et. al. (2001).

Implementation Details

A lookup table of distributions corresponding to many non-standard Gaussian distributions is constructed before simulation starts. The lookup table or database of local distributions is constructed with different Gaussian means (from approximately -3.5 to 3.5) and variances (from 0 to 2). The Z mean and variance corresponding to each distribution is calculated and saved so that the right distribution can be found during the sequential simulation procedure. This lookup table makes the CPU speed of `DSSIM-HR` virtually the same as `SGSIM` progrram or any other flavour of DSS.

The user specifies the discretization levels for the Gaussian mean, variance, and number of quantiles. Of course, as the level of discretization increases the database becomes larger and the precision with which the global distribution is reproduced increases. These distributions take very little space and this is not a practical concern.

The distribution with the closest mean and variance to the simple (co)kriging mean and variance is found in the database. It is unlikely that the distribution will have the exact mean and variance; therefore, we can either (1) interpolate in the lookup table, or (2) rescale slightly the closest distribution to have exactly the right mean and variance.

It is possible that the closest distribution is outside the generated distribution domain. This is a problem when the point- or block-scale data are inconsistent with (lower or higher than) the input global distribution used in the construction of the local distributions.

The required variogram is calculated from the original data with no normal or Gaussian transformation. The statistical input, histogram and variogram, for `DSSIM-HR` all come from original data units. The Gaussian transform is only used for to help in determining correct *shapes* for conditional distributions.

Program Details

The `DSSIM-HR` program is based on the FORTRAN 90 version of the `SGSIM` program from GSLIB. The following presents details specific to the implementation of `DSSIM-HR`.

The `bldcond` module builds the necessary database of conditional distributions to be used for the selection of distribution shapes. The input to this module includes (1) the available data set, (2) the discretization levels for the mean (nm) and variance (nv), and (3) the number of quantiles (nq) to represent each conditional distribution shape. This module evenly divides the range of Gaussian mean and variance and then back transforms each non-standard Gaussian distribution. The mean and variance of each z-conditional distribution is also calculated in original units.

The `getcond` module selects the closest local distribution from the database. The Kriging module returns with the local estimate, K_{est}, and local estimation variance, K_{std}. These values are the input to `getcond`. The routine returns a set of nq numbers in the units of the original Z variable representing the local data distribution.

The `drawcond` module draws a simulated value by Monte Carlo simulation from the nq quantiles. The exact mean and variance will not be found in the database; therefore, the simulated value is rescaled to have the exact Kriging mean and variance, that is,

$$v_{sim_{node}} = (v_{sim} - m_{sim}) \cdot \frac{K_{std}}{std_{sim}} + K_{est} \tag{5}$$

where v_{sim} is the value drawn from the unscaled distribution, m_{sim} and std_{sim} are the mean and standard deviation of the unscaled distribution, K_{est} and K_{std} are calculated correct Kriging mean and standard deviation, and $v_{sim_{node}}$ is the final simulated value. The final simulated value, $v_{sim_{node}}$, is assigned to the grid node and simulation proceeds with the next grid node.

New Parameters

Just like the GSLIB programs (Deutsch and Journel 1992), the `DSSIM-HR` program works with a parameter file. The parameters are almost exactly the same as `SGSIM` program. An example parameter file is given in Figure 1. There are two debug files for the first execution of the program. These files are for local distributions of uncertainty in data space and the mean and variance of each of those local distributions. Once these files are created for a particular input distribution there is no need to create them again; the "yes" option can be specified for the "already generated database" parameter.

The user specifies the discretization levels for mean and variance in Gaussian domain, and the number of quantiles. For the example parameter file given in Figure 1, the input values are 170, 170, and 300. This large discretization will result in very good precision in reproduction of the input global distribution. The overall simulation time, memory, and the importance of precisely reproducing the global distribution must be considered.

An important advantage of `DSSIM-HR` is that simulation is perfomed with a variogram model defined in the original data units. The input variogram should not be standardized to a sill of 1.0; the sill should be in units of the original data variance.

Example Applications

A lognormal, bimodal and uniform distribution are used in the examples shown below; see Figure 2. Arbitary variogram models were chosen. The `DSSIM-HR` program is shown

to reproduce the input histograms and variograms within expected statistical (ergodic) fluctuations.

The shape of the local distributions used for drawing the simulated value are significantly different than either the widely assumed Gaussian distribution or the original global histograms. We consider different Gaussian mean and variance values: -1.0, 0.0, and 1.0 for the mean and 0.1, 0.5, and 1.0 for the variance. The local uncertainity distributions are defined by Equation 4. The local uncertainity distributions for the lognormal distribution are shown in Figure 3. The boxed histogram corresponds to a Gaussian mean and variance of 0.0 and 1.0, which gives the original distribution. The shape of the local distributions are unique and do not look like the input lognormal distribution.

Some distributions for the bimodal distribution are shown in Figure 4. Some distributions for the uniform distribution are shown in Figure 5. In general, the shapes of local conditional distributions differ from the specified global distribution and the Gaussian distribution. The shapes systematically change from highly positively skewed to highly negatively skewed distributions. There are unique distribution shapes in the original data space for simulation. There may be distributions with different shapes that also permit reproduction of the global distribution; however, at least we know that this set will achieve our objective because it is based on the known consistent Gaussian transformation model.

Histogram Reproduction

A 2-D simulation field of 150 by 150 grid nodes is considered. An arbitrary exponential variogram model with 20% nugget effect was fixed for all cases. Other variograms were also considered and the variograms were reproduced in all cases. The actual nugget effect and variance contribution of the exponential variogram models were scaled to the global variance in each case. Four realizations were generated and the resultant histograms of simulated values are shown in Figures 6, 7 and 8.

In all the cases the realizations successfully reproduce the global mean, variance, global input histogram distribution, and variogram. There are statistical (ergodic) fluctuations. These fluctuations are an important part of uncertainty that are removed by ad-hoc post processing to enforce the global distribution.

Variogram Reproduction

Variogram reproduction is theoretically guaranteed for sequential simulation. Nevertheless, it is good practice to ensure that there are no implementation issues that artificially increase or decrease spatial correlation. The variograms were calculated for each realization and plotted with the input variograms; see Figure 9. Note the excellent agreement in all cases.

Application to Initial Potentials in Barbour County

In order to test and demonstrate the efficiency of `DSSIM-HR` under real data set, we have chosen the initial potential data set from one the field in Barbour County, West Virginia (Hohn 1997). Hydrocarbon is produced from Upper Devonian sandstones and siltstones.

The data set, taken from Hohn's book (Hohn 1997), consists of 670 wells from a 22 km by 22 km area.

A location map of the data is given in Figure 10. There are large variations with higher potential values near the middle of the study area. The global histogram of the data is also presented in Figure 10. Values up to 8500 Mscfd are seen. The global mean value is 1173.3 Mscfd and the global standard deviation is 1340.2 Mscfd.

The ominidirectional experimental variogram was calculated and fitted with the following analytical non-standard variogram model:

$$\gamma(h_1, h_2) = 772357.4 + 1023822.6\, Sph \sqrt{(\frac{h_1}{2.5})^2 + (\frac{h_2}{2.5})^2} \tag{6}$$

where Sph is shorthand notation for the common spherical variogram structure.

The study area was discretized with 150 by 150 grid nodes of size $\Delta x = \Delta y = 0.1467$ km. The database for local conditional distributions were generated. One realization for initial potentials was generated using the non-standard variogram model given above; see Figure 10. Histogram reproduction is the novel aspect of our paper; see Figure 10, which shows close reproduction of the input global histogram.

Conclusions

Sequential simulation is widely used in geostatistics. A multivariate Gaussian assumption is taken after univariate transformation to a Gaussian histogram. The advantage of working in original data units, instead of transforming to a Gaussian histogram, is straightforward integration of multiscale data. Mean, variance, and variogram reproduction is guaranteed with all implementations of sequential simulation. Nevertheless, previous efforts to work in "direct" data units have largely failed because of difficulty in reproducing the shape and details of non-Gaussian global histograms.

The Gaussian model is used to determine the conditional distribution shapes for different mean and variance values. Taken all together, these shapes ensure that the global histogram is reproduced. This leads to a direct sequential simulation `DSSIM` program with guaranteed histogram reproduction (`DSSIM-HR`). The new program was written in Fortran 90 using GSLIB conventions. Examples using synthetic and real data are shown to demonstrate the successful histogram reproduction capability of `DSSIM-HR`.

References

Caers, J., Adding local accuracy to Direct Sequential Simulation, *Mathematical Geology*, 32(7), pp.815-850, 2000a.

Caers, J., Direct Sequential Indicator Simulation, Geostats, Cape Town, 2000b.

Deutsch, C. V., Tran, T. T. and Xie, Y. L., An approach to ensure histogram reproduction in Direct Sequential Simulation, In C. V. Deutsch editor, *2001 CCG Report*, CCG, University of Alberta, Canada, 2000.

Deutsch, C. V. and Journel, A. G., *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York, 1992.

Goméz-Hernández, J. and Journel, A. G., Joint sequential simulation of multiGaussian fields, In A. Soares editor, *Geostatistics Troia 1992*, Vol. 1, pp.85-94, Kluwer, 1993.

Hohn, M. E., *Geostatistics and Petroleum Geology*, Second Edition, West Virginia, Geology Survey, 1997.

Isaaks, E. H., *The Application of Monte Carlo Method to the Analysis of Spatially Correlated Data*, Ph. D. Thesis, Stanford University, Stanford, CA, 1990.

Johnson, M., *Multivariate Statistical Simulation*, John Wiley & Sons, New York, 1987.

Journel, A. G., Class notes on sequential simulation, Stanford University, CA, 1986.

Soares, A., Oral presentation in Hedberg Conference, 2000.

Tran, T.T., Deutsch, C.V., and Xie, Y., Direct Geostatistical Simulation with Multiscale Well, Seismic, and Production Data, SPE Annual Technical Conference and Exhibition, New Orleans, September 30 - October 3, 2001, 8 pages

```
                Parameters for DSSIM-HR
****************************************
START OF PARAMETERS:
...
...
...
170 170 300                  -discretization levels for mean, variance, quantile
1                            -already generated database? (0=no, 1=yes)
l_quant_uniform_170.dbg      -file with local Z values for each quantile
l_mstdr_uniform_170.dbg      -file with local mean/std
0.50 0.29                    -global mean and variance
...
...
...
```

Figure 1: An Example Parameter file for `DSSIM-HR` Program.

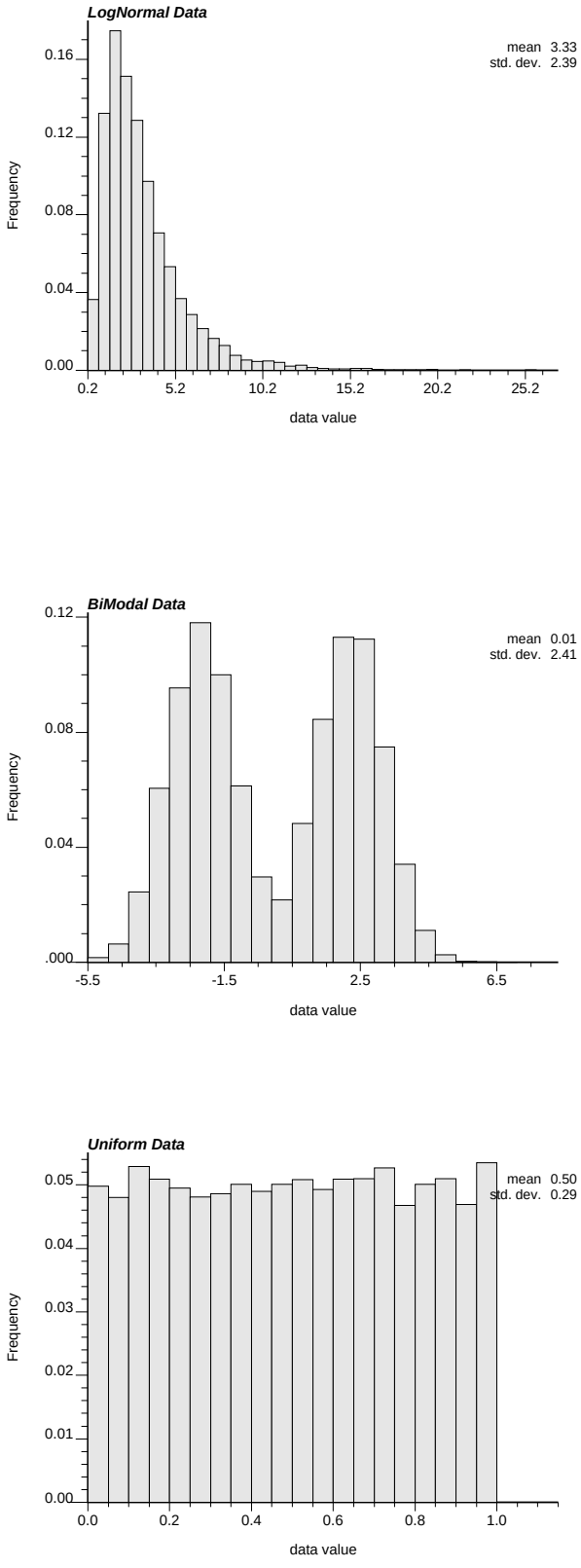


Figure 2: The three distributions used during the application of `DSSIM-HR`: lognormal, bimodal and uniform.

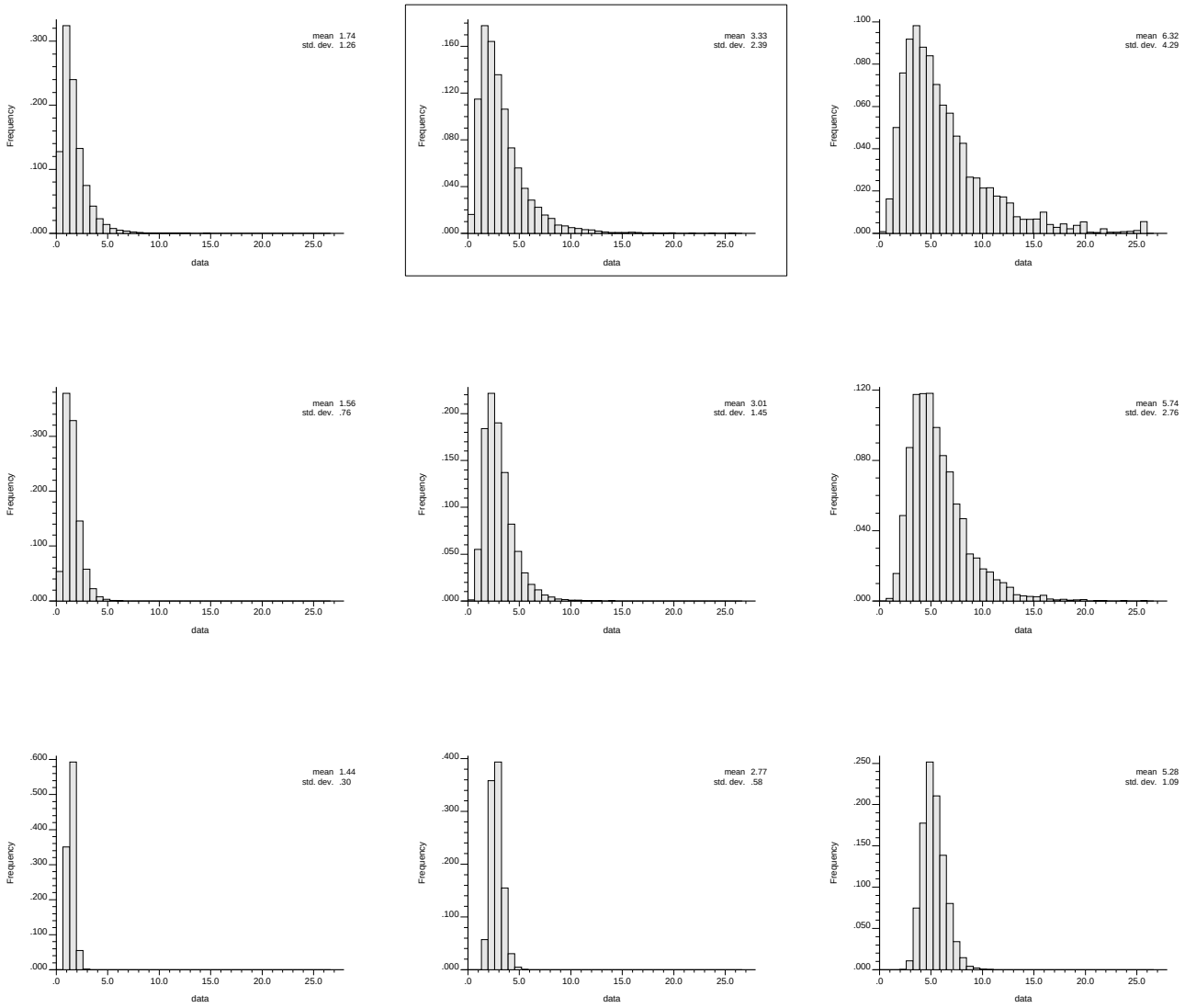


Figure 3: Local distributions for lognormal data set (from left to right: mean in normal space: -1.0/0.0/1.5; from top to bottom: variance in normal space: 1.0/0.5/0.1).

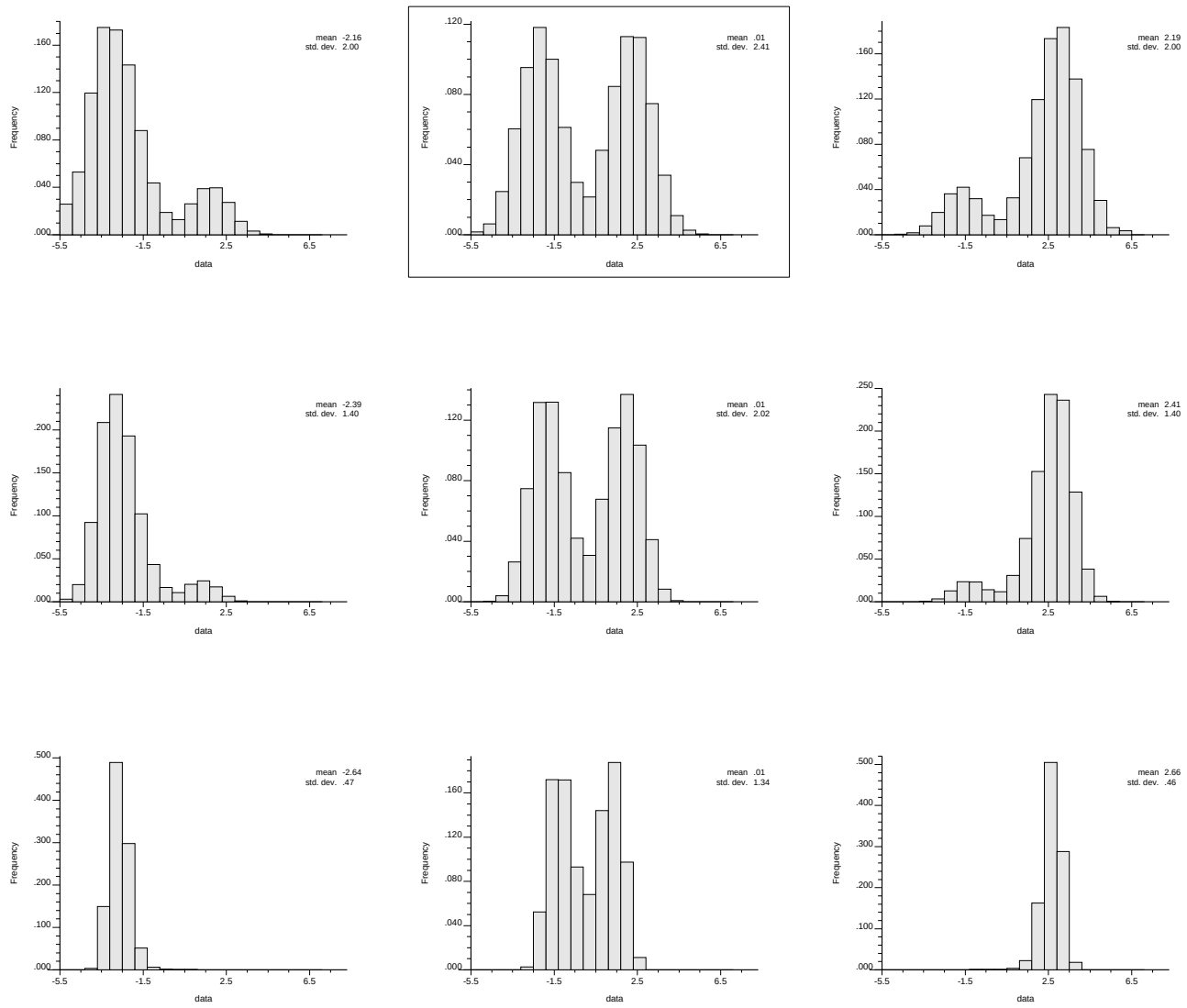


Figure 4: Local distributions for bimodal data set (from left to right: mean in normal space: -1.0/0.0/1.5; from top to bottom: variance in normal space: 1.0/0.5/0.1).

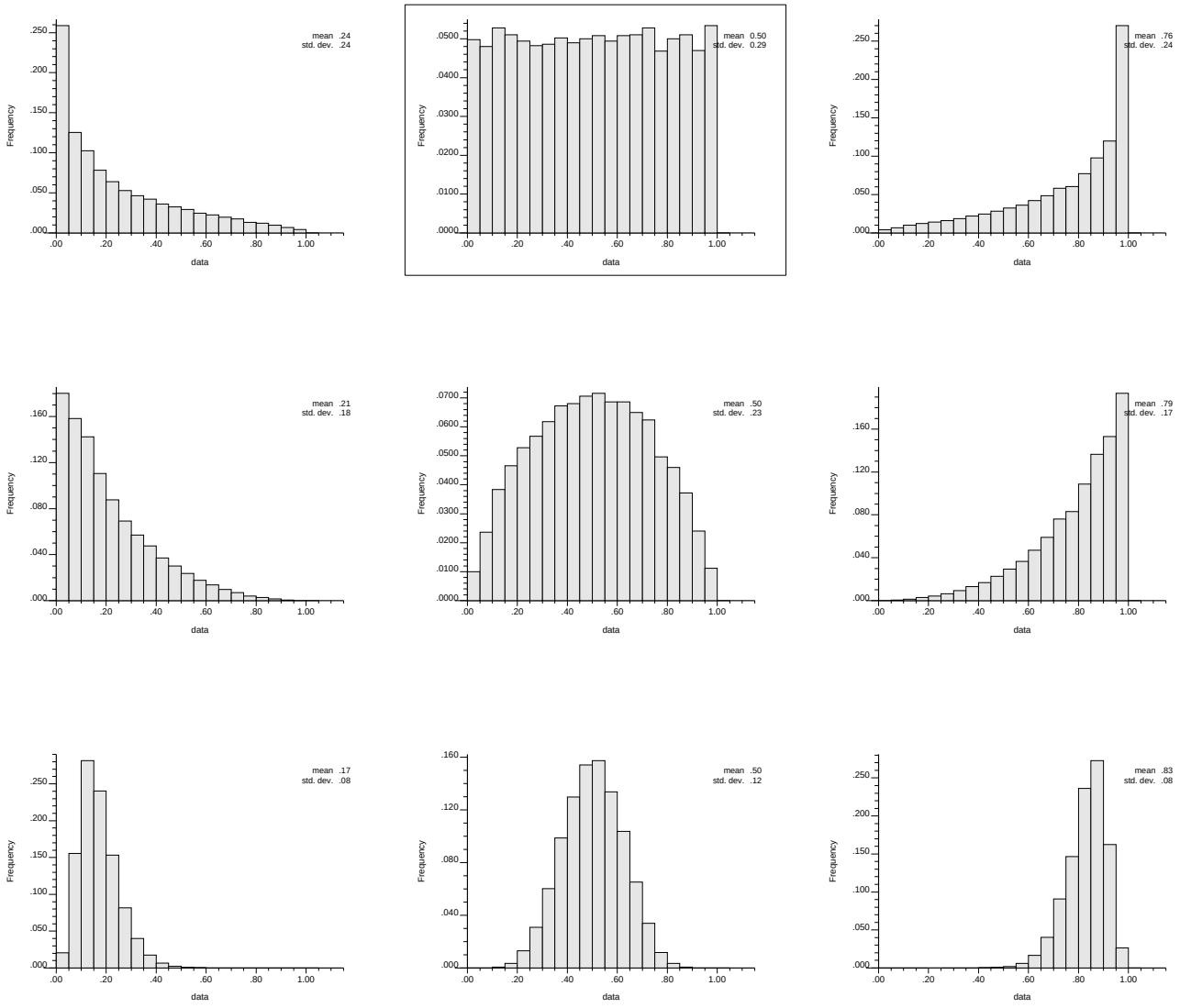


Figure 5: Local distributions for uniform data set (from left to right: mean in normal space: -1.0/0.0/1.5; from top to bottom: variance in normal space: 1.0/0.5/0.1).

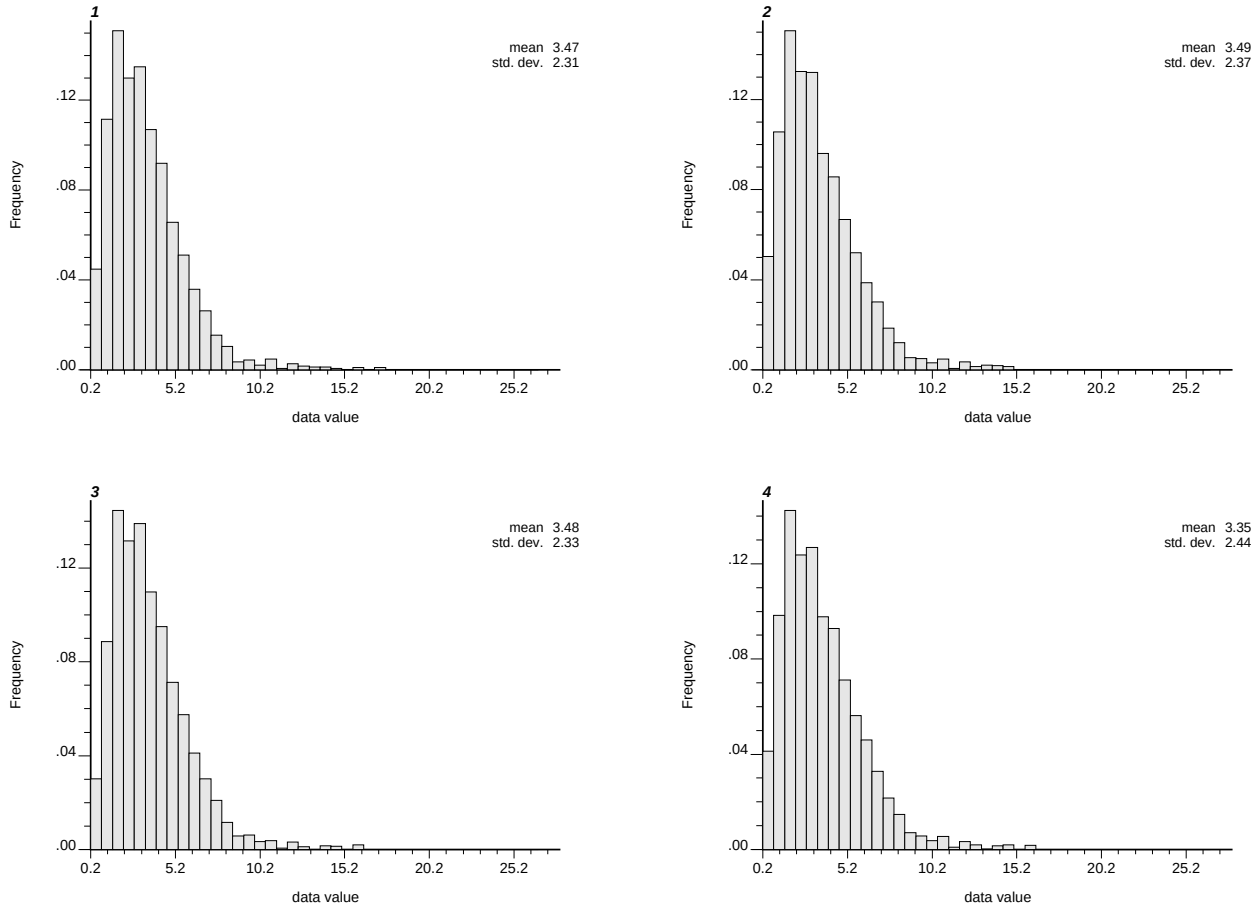


Figure 6: Reprodution of the histograms for four different realizations generated by `DSSIM-HR` using the input lognormal distribution.

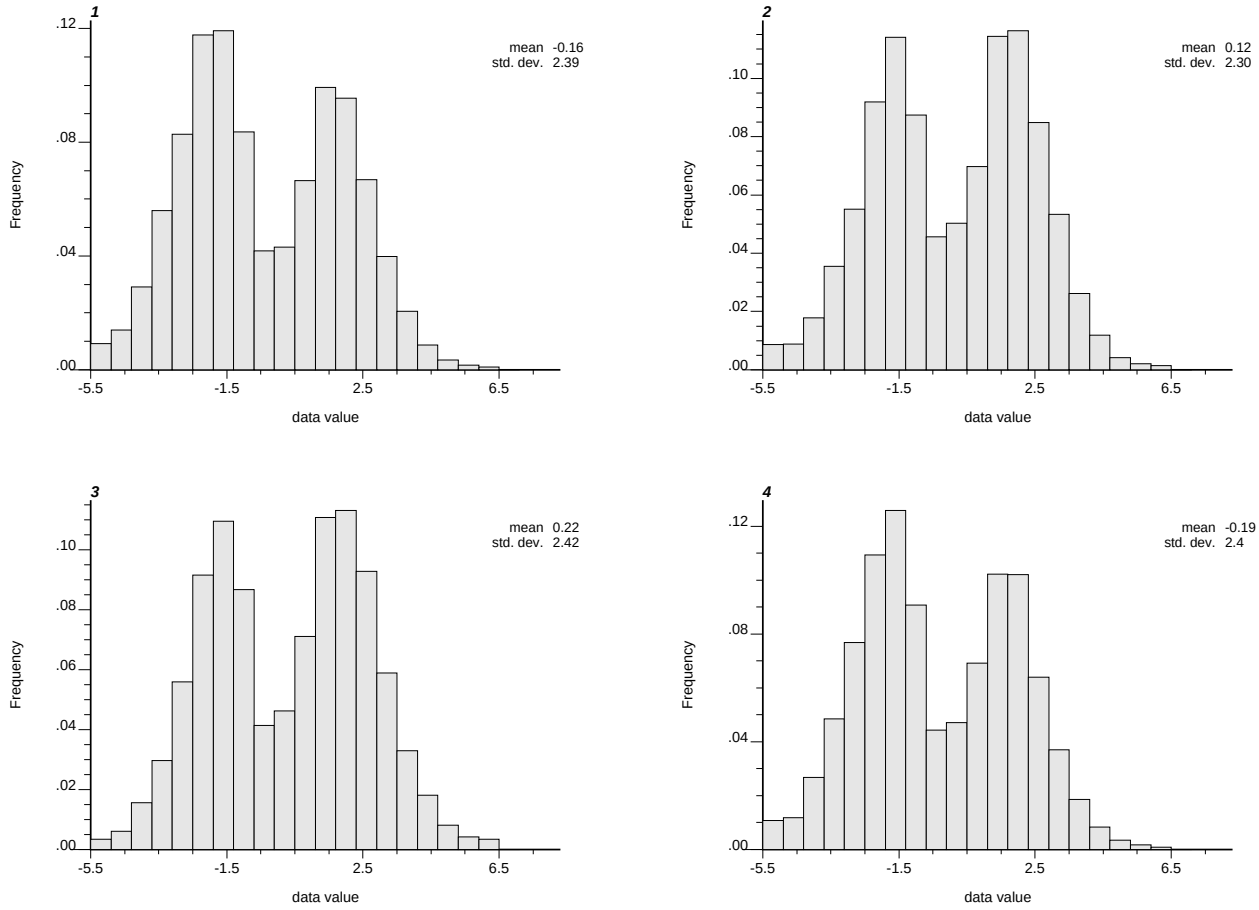


Figure 7: Reprodution of the histograms for four different realizations generated by `DSSIM-HR` using the input bimodal distribution.

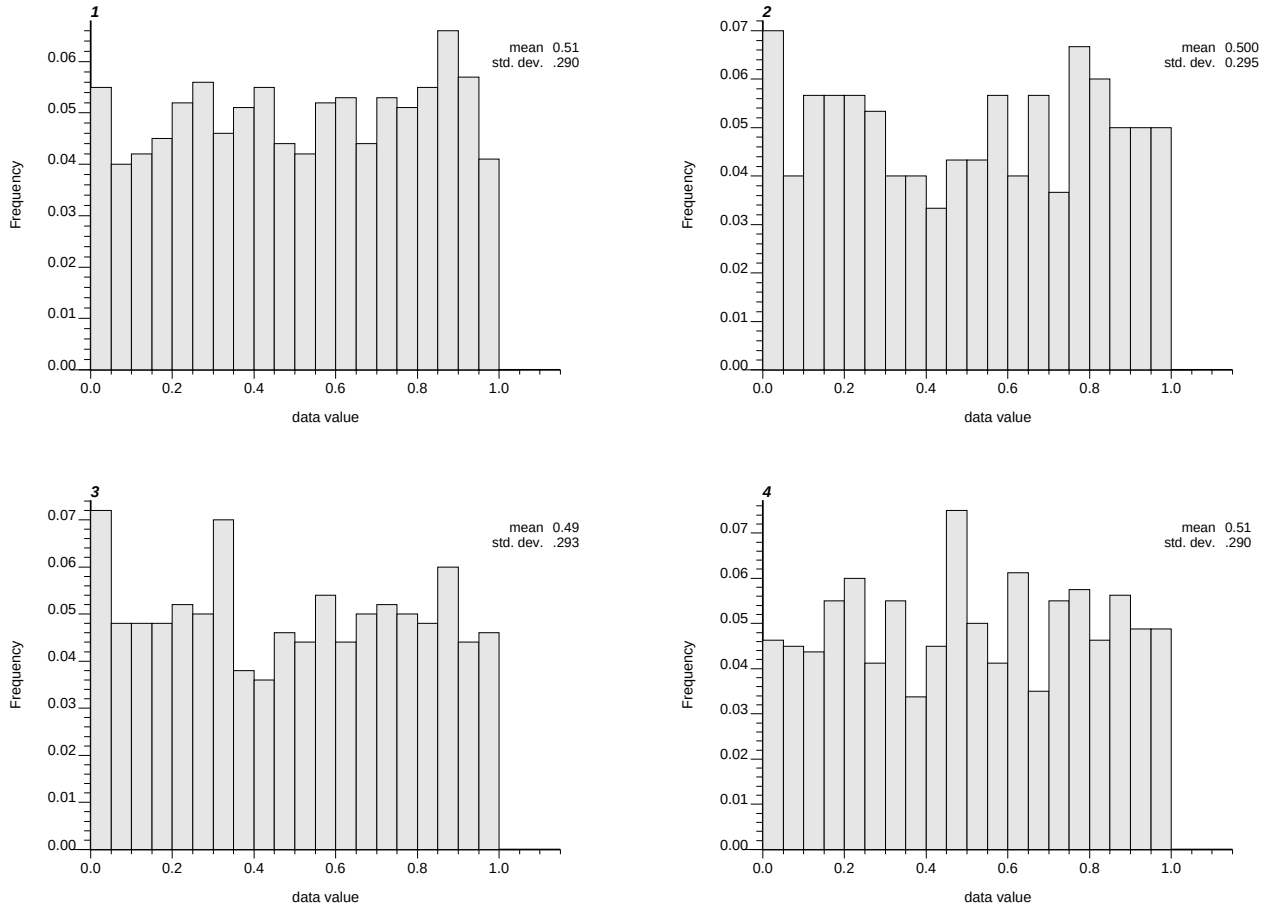
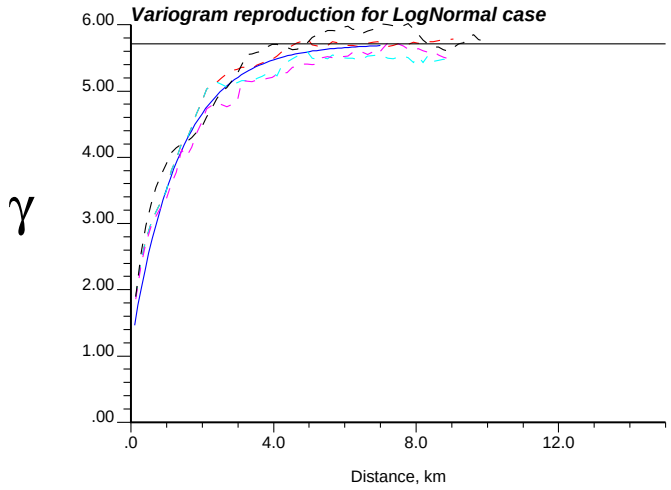
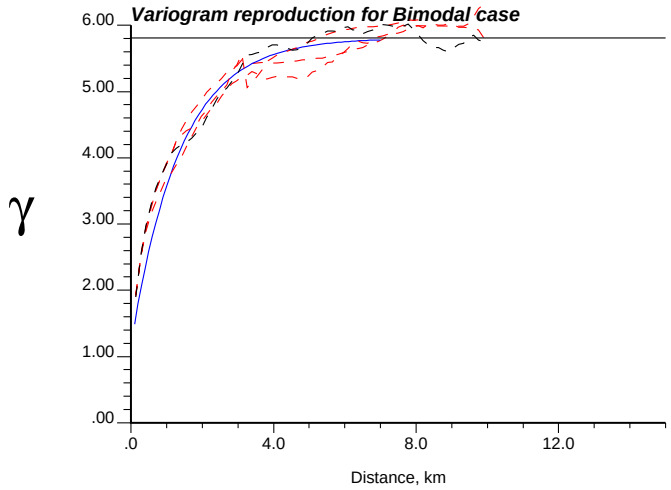


Figure 8: Reprodution of the histograms for four different realizations generated by `DSSIM-HR` using the input uniform distribution.





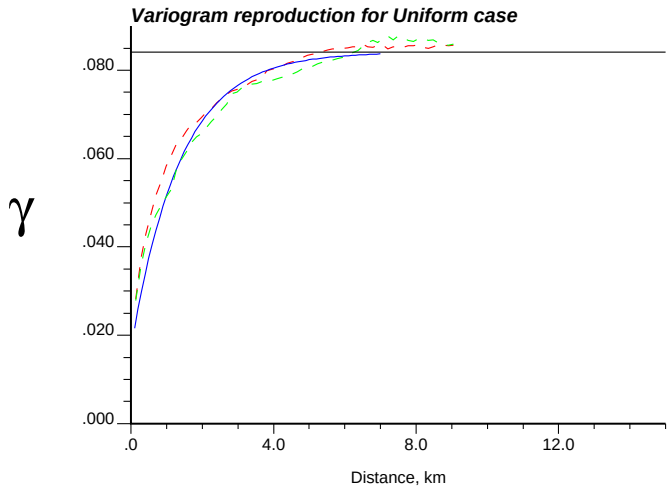


Figure 9: Reprodution of the Variogram for lognormal, bimodal, and uniform distribution.

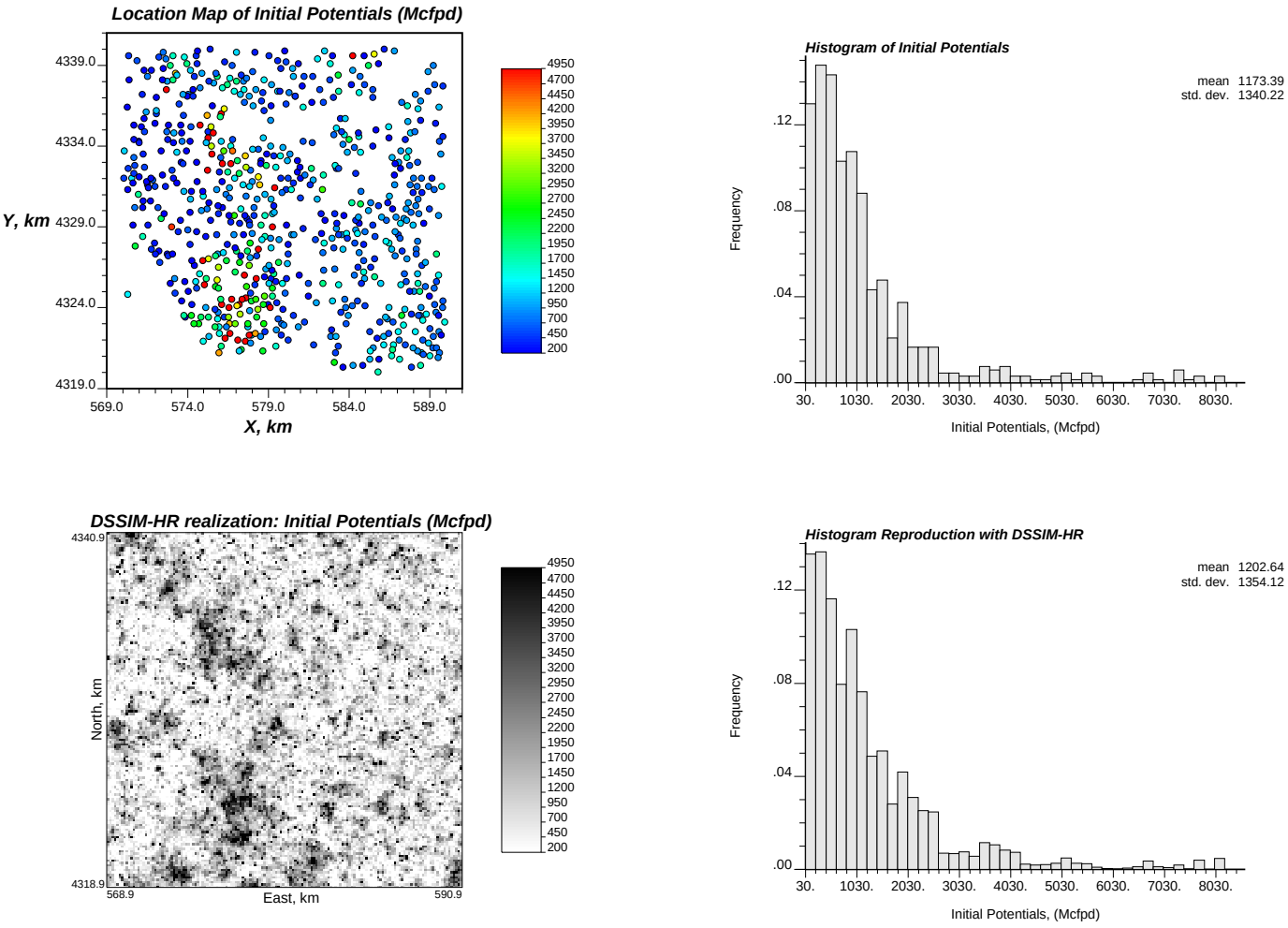


Figure 10: Location map for initial potentials (top right), original input histogram (top left), map of `DSSIM-HR` realization (bottom right), and histogram of the simulated inital potentials with `DSSIM-HR` (bottom left).