A Preliminary Report on:

An Approach to Ensure Histogram Reproduction in Direct Sequential Simulation

C. V. Deutsch
Centre for Computational Geostatistics
University of Alberta, Edmonton, Alberta

T. T. Tran
Chevron Petroleum Technology Company
San Ramon, California

Y-L. Xie
Pacific Northwest Laboratories
Richland Washington

Abstract

There are many advantages to direct sequential simulation (DSS). Avoiding a Gaussian or indicator transform permits *correct* integration of data of different volume supports; the non-linear transform of Gaussian and indicator simulation techniques only allows reproduction of the *trend* or *rank* of the data at different scale than the simulated grid blocks. In particular, simulation of unstructured grids requires consideration of previously simulated grid blocks at many different scales. Moreover, DSS allows a more correct approach to account for secondary data that is often at larger scale.

The main limitation of DSS has been difficulty with histogram reproduction, that is, the global histogram of the simulated values ends up as a hybrid of the data histogram and a symmetric Gaussian distribution due to the central limit theorem. Attempts to honor the global histogram *by-construction* have largely failed. A post-processing quantile transformation is required to enforce the global histogram. This post-processing removes ergodic fluctuations and destroys reproduction of large-scale data.

We introduce a procedure that allows the global histogram to be reproduced, within ergodic fluctuations, by-construction in a theoretically correct manner. The mean and variance of the conditional distribution at each step of DSS come from simple (co)kriging (as it must). The shape of the conditional distribution is taken from the correct shape if a univariate Gaussian transform were used. The method does *not* assume multiGaussianity nor are the data transformed to a Gaussian distribution; the shape implicit to the Gaussian model is simply used as a viable shape to honor all input data and the global distribution. The approach and implementation details are explained.

Introduction

The sequential paradigm to simulation has become increasingly popular. It has advantages over classical geostatistical techniques such as matrix methods, spectral methods, and moving average methods. These classical methods only work with multivariate Gaussian distributions and require a kriging step to make the simulated realizations honor local data. The sequential simulation

approach is somewhat more flexible for continuous and categorical variables and accomplishes the simulation in one step.

Sequential simulation methods have historically been applied to transformed variables, that is, a Gaussian transform of continuous variables and an indicator transform for categorical variables. The indicator transform could also be used for continuous variables; however, this approach is more demanding for inference – we do not concern ourselves with indicator methods in this report. The application of Monte Carlo simulation from a series of conditional distribution is a classical statistical procedure that is well grounded in Bayesian statistics. 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 sequential paradigm is not approximate; however, care must be taken to avoid artifacts by poor implementation decisions such as using too few previously simulated values.

The Gaussian transformation makes implementation of sequential simulation remarkably straightforward. A decision is made to model the full multivariate distribution with a Gaussian distribution after univariate transformation to a normal or Gaussian distribution. Then, the conditional distributions at each step of the sequential simulation are Gaussian in shape with mean and variance given by simple (co)kriging. The original Z variable is transformed to a Y Gaussian variable, simulation is done in Y Gaussian space, and the simulated y values are back transformed to z values. The covariance or variogram of the Y random variable is correct.

Variogram reproduction is guaranteed by use of all data and previously simulated grid blocks and by application of simple kriging. In practice, a limited search neighborhood is used, but variogram reproduction can be checked and more data used if variogram reproduction is deemed unacceptable. 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. Sequential Gaussian simulation (SGS) is arguably the most powerful and commonly used geostatistical simulation technique at the present time.

The histogram of any particular SGS realization does not match the input histogram exactly. The back transformation in SGS would only impose the histogram exactly if the Gaussian or normal values were exactly normal with a mean of 0, variance of 1, and correct shape. Simulated realizations show statistical or ergodic fluctuations between realizations. These variations are an important part of uncertainty; we will expect variability in the sample statistics over any study area of finite size. It is wrong to transform the results of SGS to impose the histogram exactly.

Working in Gaussian *space* makes calculations straightforward; however, it was shown early in the development of sequential techniques that the variogram structure is reproduced without transformation to Gaussian space (Journel, class notes, 1987). Direct sequential simulation (DSS), applied directly with the original Z data values, would lead to simulated values that follow the correct variogram. The Monte Carlo simulation at each step must consider probability distributions with the mean and variance given by simple (co)kriging, but the shape of the conditional distribution does not matter if we only want to reproduce the mean and variogram. Until now, there has been no good way to decide what shape of distribution to use in DSS. In general, regardless of the shape chosen for the conditional distributions, the global histogram of the final values taken altogether is *not* reproduced. The histogram is important; it is a first order statistic that has a first order affect on calculations made with the simulated realizations. The inability of DSS to honor the input histogram has been a significant problem.

Notwithstanding this significant problem with DSS, interest in a *direct* method has grown. The main reason is that we must use a direct method to simultaneously account for data of different volumetric scale. Transforming data of different scale to Gaussian space is problematic: the transform to a Gaussian distribution is non-linear and yet most averaging is linear (porosity) or

very particular (permeability). A direct method would avoid the need for this problematic transformation. There are other reasons such as the integration of secondary variables at the correct scale and with the correct level of precision. The problem of global histogram reproduction must be addressed for successful application of DSS.

The same quantile-transformation procedure used to transform original Z values to Gaussian Y values can be used to transform the output-simulated values from direct simulation to the correct input histogram. The problem with this back transformation is that the final global histogram has no uncertainty (ergodic fluctuations) and, more importantly, large scale data is not reproduced. The transformation can be modified so that local hard data are reproduced (the values before and after transformation can be averaged together with a special weighting function); however, the problems of block data statistical fluctuations are important.

Caers (CapeTown, 2000) 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 reproduces the global histogram. This approach also removes most ergodic fluctuations and could introduce artifacts.

Soares (Hedberg Conference, 2000) proposed a different approach to reproduce the histogram in DSS. The central idea of Soares's proposal was also to draw values selectively based on the kriged mean and variance. The procedure does not seem to work well except when the variogram is nearly pure nugget effect.

We propose another method for histogram reproduction. The challenge has always been to determine the shape of the conditional distribution. The true beauty of the Gaussian approach is that the shape is always Gaussian or normal. In original Z units, there is no way to know the right shape so that the final simulated values reproduce the global histogram when taken altogether. The key ideas behind our method is to (1) work in original Z space, that is, a true DSS application, and (2) work out the shape of the conditional distributions as a function of their mean and variance using the normal-score or Gaussian transformation. We can have the best of both worlds, that is, no data transformation and guaranteed reproduction of the input histogram within statistical fluctuations.

The method will be described in detail with attention to practical implementation details and limitations.

A Look at Sequential Simulation

Sequential simulation is described in many sources including Deutsch and Journel, 1998 and Goovaerts, 1997; the details will not be repeated here. The properties of kriging and expected values recalled below give one view of sequential simulation. There are, of course, different ways of deriving and explaining the theoretical basis for sequential simulation. Nevertheless, it is correct to say that there are two key results that makes sequential simulation work: (1) the covariance reproduction property of kriging, that is, the covariance between kriged values and the original data values follows the input model, and (2) addition of an independent random variable to a kriged estimate increases variance without changing the covariance. The combination of these two results makes sequential simulation work. Following are informal proofs of each result.

Proof of the covariance reproduction property of kriging: first consider the simple kriging estimate, system of equations, and kriging variance for a standardized variable y:

$$
\begin{aligned}
y^*(\mathbf{u}) &= \sum_{i=1}^{n} \lambda_i \cdot y(\mathbf{u}_i) \\
\sum_{j=1}^{n} \lambda_j \cdot C(\mathbf{u}_j - \mathbf{u}_i) &= C(\mathbf{u} - \mathbf{u}_i), \quad i = 1, \ldots, n \\
\sigma_K^2(\mathbf{u}) &= \sigma^2 - \sum_{i=1}^{n} \lambda_i \cdot C(\mathbf{u} - \mathbf{u}_i)
\end{aligned} \tag{1}
$$

Standard geostatistical notation is used, that is, **u** represents a 3-D location vector, the subscript $i=1,\ldots,n$ relates to the available data, and $C(\mathbf{h}=\mathbf{u}_1-\mathbf{u}_2)$ is the stationary covariance $(1-\gamma(\mathbf{h}))$. The covariance between the kriging estimate $y^*(\mathbf{u})$ and one particular data value $y(\mathbf{u}_i)$ can be calculated:

$$
\begin{aligned}
Cov\left(y^*(\mathbf{u}), y(\mathbf{u}_i)\right) &= E\left\{y^*(\mathbf{u}) \cdot y(\mathbf{u}_i)\right\} \\
&= E\left\{\left(\sum_{j=1}^{n} \lambda_j \cdot y(\mathbf{u}_j)\right) \cdot y(\mathbf{u}_i)\right\} \\
&= \sum_{j=1}^{n} \lambda_j E\left\{y(\mathbf{u}_j) \cdot y(\mathbf{u}_i)\right\} \\
&= \sum_{j=1}^{n} \lambda_j \cdot C(\mathbf{u}_j - \mathbf{u}_i) \\
&= C(\mathbf{u} - \mathbf{u}_i)
\end{aligned} \tag{2}
$$

Note that (1) considering a standardized variable removes the need to consider the product of means in the covariance, and (2) the final substitution comes from the simple kriging equations (1). This is a powerful result. The covariance between the kriging estimate $y^*(\mathbf{u})$ and each of the data values $y(\mathbf{u}_i)$, $i=1,\ldots,n$ using in kriging is correct. This could be used as justification for the use of kriging and the sequential simulation as a way to enforce covariance reproduction between all simulated values.

Although sequential kriging would lead to values that reproduce the required covariance, the resulting values would be too smooth, that is, the variance would be too small. Although the covariance is correct, the variogram is not because the variance is too small. Another very important property of kriging is that the amount of smoothness can be calculated ahead of time. The smoothing is exactly the simple kriging variance. The variance of the kriged estimate is the stationary variance minus the kriging variance:

$$
Var\left\{y^*(\mathbf{u})\right\} = \sigma^2 - \sigma_k^2(\mathbf{u}) \tag{3}
$$

This leads to the second key aspect of sequential simulation. The variance of the estimates must be increased by the kriging variance. A random residual is added at each step:

$$
y^s(\mathbf{u}) = y^*(\mathbf{u}) + r(\mathbf{u}) \tag{4}
$$

The random residual is drawn by Monte Carlo simulation (independently) from a distribution with zero mean and variance equal to the kriging variance $\sigma_k^2(\mathbf{u})$. Regardless of the distribution for $r(\mathbf{u})$, the expected value of the simulated value $y^s(\mathbf{u})$ is that of the kriged value:

$$
\begin{aligned}
E\left\{y^s(\mathbf{u})\right\} &= E\left\{y^*(\mathbf{u}) + r(\mathbf{u})\right\} \\
&= y^*(\mathbf{u}) + 0 \\
&= y^*(\mathbf{u})
\end{aligned} \tag{5}
$$

Moreover, the variance of the simulated value has been restituted to the full variance σ^2, which is required for the stationary random function:

$$
\begin{aligned}
Var\{y^s(\mathbf{u})\} &= Var\{y^*(\mathbf{u}) + r(\mathbf{u})\} \\
&= \sigma^2 - \sigma_k^2(\mathbf{u}) + \sigma_k^2(\mathbf{u}) \\
&= \sigma^2
\end{aligned} \tag{6}
$$

The most important result of adding a random number $r(\mathbf{u})$ with no spatial correlation is that the covariance reproduction property of kriging (2) is not changed:

$$
\begin{aligned}
Cov\left(y^s(\mathbf{u}), y(\mathbf{u}_i)\right) &= E\left\{\left(y^*(\mathbf{u}) + r(\mathbf{u})\right) \cdot y(\mathbf{u}_i)\right\} \\
&= E\left\{\left(\sum_{j=1}^{n} \lambda_j \cdot y(\mathbf{u}_j)\right) \cdot y(\mathbf{u}_i) + r(\mathbf{u}) \cdot y(\mathbf{u}_i)\right\} \\
&= \sum_{j=1}^{n} \lambda_j E\left\{y(\mathbf{u}_j) \cdot y(\mathbf{u}_i)\right\} + 0 \\
&= \sum_{j=1}^{n} \lambda_j \cdot C(\mathbf{u}_j - \mathbf{u}_i) \\
&= C(\mathbf{u} - \mathbf{u}_i)
\end{aligned} \tag{7}
$$

The expected value of the random number $r(\mathbf{u})$ multiplied by data value at i $\underline{y}(\mathbf{u}_i)$ is zero because they are independent of each other and the expected value of $r(\mathbf{u})$ is zero. Independence entails that the expected value of the product is the product of expected values (Bayes Theorem); hence the covariance is correct.

Drawing a random number $r(\mathbf{u})$ from a distribution with a mean of zero and a variance of $\sigma_k^2(\mathbf{u})$ and adding the kriged estimate $y^*(\mathbf{u})$ is identical to drawing from a distribution with a mean of $y^*(\mathbf{u})$ and a variance of $\sigma_k^2(\mathbf{u})$. As mentioned in the introduction, the issue comes down to what shape of distribution should we use for this distribution. Kriging can be thought of as an additive or averaging procedure. The central limit theorem tells us that the average of independent (which the $r(\mathbf{u})$ values are) identically distributed values tends toward a Gaussian distribution; therefore, if the original y data are standard Gaussian and a Gaussian shape is used for the $r(\mathbf{u})$ values, the final result will tend toward the correct standard Gaussian histogram. A more formal proof of this could be offered; however, geostatisticians know that sequential Gaussian simulation (SGS) works in practice. Data are transformed to a standard Gaussian distribution, sequential simulation proceeds, and the simulated values are back transformed afterwards.

Gaussian simulation works fine when the data are all at the same volumetric scale and the simulation is being conducted at that exact same scale. Gaussian simulation does not work in presence of multiple data at different scale or when we want to simulate at a different scale from the data. To use Gaussian simulation, we must assume that the variable averages linearly after Gaussian transformation, which is not the case. We would like to work in original Z data units for variables that average linearly and, perhaps, in ω-power law transformed space for permeability that approximately follows a constant power-law average. Working in original data units or with some arbitrary non-Gaussian transform requires us to address the question of *what shape of distribution do we use for r in sequential simulation?*

Failure to address this question satisfactorily will result in simulated realizations that reproduce (1) the conditioning data at the scale of the simulation, and (2) the variogram. The simulated realizations will not, however, reproduce the global histogram of the variable.

Methodology for Histogram Reproduction

The *correct* shape of the conditional distributions in sequential simulation is known for the Gaussian case because we have a model for the full multivariate distribution. An evident approach to work in original Z units is to derive an alternative multivariate model. There are significant problems with this idea: (1) there are no alternative models as tractable as the Gaussian model, and (2) a non-Gaussian analytical model would necessarily follow some other parametric model that would not likely match our data.

Other ideas have been put forward. Caers suggests dynamic monitoring of the global distribution and using selective sampling to ensure the global is approximately reproduced. Soares proposes a different selective sampling procedure where values are drawn from particular regions of the global distribution depending on the kriging mean and variance. These techniques do not ensure the global distribution is reproduced and remove important fluctuations in the result.

Rather than attempt selective sampling that destroys the statistical properties of our final models, we propose to use a family of distribution shapes that we infer from the multivariate Gaussian transform procedure.

Conditional Distribution Shapes from Multivariate Gaussian Model

Consider an original Z variable with stationary histogram $F_Z(z)$. In the Gaussian approach this variable is transformed to a Y variable with stationary standard normal distribution $G(y)$. The quantile or normal-score transformation is widely used to transform any z-value to a corresponding y-value:

$$y = G^{-1}\left(F_Z(z)\right) \tag{8}$$

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

$$z = F_Z^{-1}\left(G(y)\right) \tag{9}$$

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

A fantastic property of the multivariate Gaussian model is that we know the shape of every conditional distribution: univariate Gaussian! The mean and variance are given by kriging. The distribution of uncertainty in Z space can be determined from the non-standard univariate Gaussian distribution by Monte Carlo simulation (drawing L random y values) or straightforward back transform of L regularly spaced quantiles:

$$z^l = F_Z^{-1}\left(G\left(\frac{G^{-1}(p^l) - y^*}{\sigma_k}\right)\right), l = 1,\ldots,L \tag{10}$$

where y^* and σ_k are the mean and variance of the non-standard Gaussian distribution of uncertainty, and the $p_l, l=1,\ldots,L$ values are uniformly distributed between 0 and 1. The distribution of uncertainty in Z space is assembled from the z^l, $1,\ldots,L$ values. There is no analytical expression for this distribution, aside from expression 10; nevertheless, the distribution is completely defined:

$$F_{Z,y^*,\sigma_k}(z) \tag{11}$$

We add the Gaussian parameters as subscripts to denote a conditional distribution relating to a particular conditional distribution in Gaussian space. The shape, mean, and variance of this

distribution depend on the original Z distribution, but are not the same as the original Z distribution.

Figure 1 shows this concept graphically. The top row shows histograms of "real" Z space and Gaussian Y space. The second row shows the cumulative Z and Y distribution functions, $F_Z(z)$ and $G(y)$. This is the only way to go between Z and Y units. The lower two rows of figures shows histograms and cumulative distribution functions of conditional distributions (low and high mean / low and high variance). 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. Our proposal is to use those known shapes in DSS.

DSS with Predetermined Conditional Distribution Shapes

Our proposal is for direct sequential simulation (DSS), that is, all kriging and simulation is performed in original Z variable units (or with an appropriate ω-power law transform). We only use the Gaussian transform to get the *shape* of the conditional distributions. In concept, our proposal consists of conventional sequential simulation with the following modifications:

1. Determine the appropriate mean and variance in Z units by (co)kriging using all relevant original data and previously simulated grid nodes or blocks, $z^*(\mathbf{u}) / \sigma_z^2(\mathbf{u})$.
2. Find the corresponding Gaussian mean and variance $y^*(\mathbf{u}) / \sigma_y^2(\mathbf{u})$ that would yield a z-conditional distribution with the z mean and variance from step one ($z^*(\mathbf{u}) / \sigma_z^2(\mathbf{u})$).
3. Draw a simulated z value from this conditional distribution, that is, $F_{Z,y^*(\mathbf{u}),\sigma_k(\mathbf{u})}(z)$, see relations (10) and (11).

Step 2 in this procedure could potentially require significant computing effort; however, for practical implementation, we build a database of local distributions with different y means and variances. Determining the correct local distribution shape amounts to a fast table look-up.

Theoretical Justification

Clearly this proposal will create realizations that reproduce the (1) local data at point and block scale since kriging is done in original Z data units, and (2) the mean and variogram of the Z variable because of the principles of DSS simulation described above. It remains to be shown that the global distribution of the Z variable, $F_Z(z)$, is reproduced.

There is an appealing argument to be made about the parallel between Z space and Y space and the linkage of equations (8) and (9). A p-quantile in Y space is directly linked to the corresponding p-quantile in Z space and. Simulating a y value in normal space and back-transforming to z space afterwards is identical to drawing the same uniform random number and directly generating a z value from the corresponding distribution (11).

This is not a proof. We must demonstrate the link between $y^*(\mathbf{u}) / \sigma_y^2(\mathbf{u})$ and $z^*(\mathbf{u}) / \sigma_z^2(\mathbf{u})$ using the same data and the correct y and z covariance structure.

The procedure works in practice.

Examples

Four data sets were used to test the algorithm. Data set one contains permeability data from two vertical wells. The horizontal separation of the two wells is 600 meters and the vertical spans about 100 meters. Data set two is a 3D data set of copper grades. Data sets three and four are synthetic data sets with nearly Gaussian histograms. The histograms for the four datasets are shown on Figure 2.

The experiment variograms of the first two sets were calculated and modeled. Figure 3 shows the experimental and model variogram for the first two datasets. The variogram models for all four datasets are given below:

Data Set	*Mean/Std*	*Variogram Model*
1	1176/1141	$\Gamma = 100000 + 600000Sph\begin{pmatrix} 4 \\ 200 \end{pmatrix} + 601884Sph\begin{pmatrix} 13 \\ 600 \end{pmatrix}$
2	1.02/0.52	$\Gamma = 0.1 + 0.03Sph\begin{pmatrix} 1 \\ 100 \end{pmatrix} + 0.1404Sph\begin{pmatrix} 100 \\ 100 \end{pmatrix}$
3	15.18/5.11	$\Gamma = 23.5009 + 2.6112Sph(200)$
4	30.0/5.0	$\Gamma = 2.5 + 22.5Sph\begin{pmatrix} 10 \\ 10 \end{pmatrix}$

Since there is no direct control on the mean and standard deviation of the local distribution in the original data space, the distribution domain of local distributions in the original space is investigated. 100,000 random pair of mean and standard deviation are generated in the Gaussian space. The mean obeys the standard normal between –3.5 to 3.5 and the standard deviation is a uniform distribution between 0 and 1. Each Gaussian distribution corresponding to a mean/standard deviation pair is back transformed to the original data space, and the mean and standard deviation in the data space are calculated.

Figures 4 show the scatter plots (red dots) of mean versus standard deviation in the original data space. The cross of two blue lines corresponds to the global distribution. The shape of such domains depends on the data distributions. It seems the closer to the Gaussian of the data distribution, the more symmetrical of the domain. Also it seems the shape of such a domain is very sensitive to the shape of the data distribution. For example, both Data set III and IV are Gaussian, but data set IV seems like a little bit more skewed, the distribution domain in the original data space is more asymmetrical.

In the practice, the storage and retrieval of 100,000 distributions will be inappropriate and limited distributions will be generated as an approximation of that domain. Usually regular spacing point in the mean and variance axes are taking and distribution are generated based on these mean/variance pairs. The green dots in the Figures are scatter plots of the mean versus standard deviation of the distributions in the database when discretizing the mean and variance axes with 101 levels. Figure 5 shows the approximation when discretizing the domain using 11 levels. It is obviously that the more discretization levels, the better the approximation of the domain. However, it should compromise in practice between the precision of the approximation and the memory and computation requirement.

Direct simulations were carried out for each data set based on 101 discretization levels for both mean and standard deviation. The simulation domains are listed below.

Data Set	*Simulation Domain*
1	120×550 with $\Delta x = 5.0$, $\Delta y = 0.2$
2	400×600 with $\Delta x = 5.0$, $\Delta y = 5.0$
3	10000 with $\Delta x = 0.1$
4	250×250 with $\Delta x = 0.2$, $\Delta y = 0.2$

The blue dots shown in Figure 4 are the scatter plot of Kriging mean versus Kriging standard deviations. It is noticed that for the two non-Gaussian data sets the domain of the local distributions generated does not cover the real domain from Kriging. That poses a serious problem because one cannot find a close distribution for those blue dots located outside the distribution domain. The adoption of the available local distribution with very different mean/standard deviation will change the distribution of the simulated value very much.

In order to alleviate this problem, the domain of the local distributions is increased in the creation. Although we know variance 1 in the normal space represent the variance of the global distribution which has the maximum variance, we increase this limit to 2 with the hope that we will get a wider coverage of the distributions in the original data space. Figure 6 show the enlarged distribution domains for all the data sets together with the actual Kriging space.

From the figure, it is noticed that the doubling the variance in the Gaussian space does not have the same significance to the variance change in the original data space. Although the alleviation of the non-coverage problem, it does not eliminate it.

The proposed procedure aims to get right shape of the local distributions. Figures 7 to 10 show several local distributions generated for the four data sets.

The local distributions appear systematical change from highly positive skewed to highly negative skewed through the transition of the global distribution.

Figure 11 shows the histograms of the simulated values for four data sets. Comparing to the data histograms shown in Figure 2, the histogram reproduction for the two Gaussian data sets (data set III and IV) are quite good, but the reproduction of the other data sets are not so satisfactory, especially for data set I. This is not beyond the expectation given the non-covered Kriging space by the domain of the local distributions shown in Figures 4 and 6. By taking distribution inside the domain as a replacement for those Kriging dots lie outside the domain shifts the values systematically to a higher ones, which may explain why the quartiles of the simulated values are all bigger than that in original data (but why the renormalization by the Kriging mean/standard deviation does not avoid this, does this suggest the distribution shapes of those outside dots are significantly different from the ones selected from the distribution database??)

From Figure 6 those Kriging estimates outside the domain of the local distributions have low Kriging mean (even lower than data minimum) but still have high enough variance. Theoretically, the further away of the Kriging mean away from global mean, the more known information are used in such a determination, the less the Kriging variance should be. In Kriging, the Kriging variance is determined independent from data values as well as the weights, the Kriging estimate is determined subsequently. We have checked the weights for some Kriging dots outside the distribution domain, and we do not find obvious weird weights associated. Mostly such situations associate with the appearance of negative weights. A small negative weight associated with a large data value may lead to a Kriging mean far less the minimum data value. However sometimes a set of descent positive weights may also lead to very small Kriging mean because

simple Kriging is used to kriging the residual from the global mean. Even for a non-problematic Kriging mean, the inability to find a close local distribution from the distribution database, the subsequent rescaling of the simulated value may also lie way out of the data range. Even though the local distribution used for drawing the simulation is well behaved within the data range, the rescaling of the simulated value based on the Kriging mean/standard deviation will cast the simulated value far way from the data range.

If no constraints applied to the simulated value, the accumulation of weird simulated values (which will affect subsequent Kriging mean) will result much more Kriging dot outside the distribution domain. The blue dots plots shown in Figure 6 actually are after constraining the simulated values inside the data range. The constraints of simulated value inside the data range should not be a good option although it alleviates the phenomenon of Kriging dots outside of the distribution domain. In sgsim, there is no such a constraint. Instead, the tail extrapolation option allows the simulation has simulated values outside data range but within the low/up data limits. Since in sgsim, the Kriging is on the nscore space and the simulated value will be back transformed into data space, it is unlikely to have weird simulated value. In direct simulation, the Kriging is on the data space, and there is no control at all which kind of Kriging estimate occurs. Applying constraints to the Kriging mean may violate the principle behind, but no constraints may results more deteriorated results.

However, the reproduction of the histograms are not so bad even though it seems there are lots of Kriging dots are outside of the local distribution domain. I guess, the significance of the influence of those outside dots depends on the difference of the shapes in the local distributions of those dots from the local distribution used in the simulation. If no difference in the shape, it actually does not matter because the simulated value will be rescaled based on Kriging estimates anyway. However, there is no way to compare the distributions because we do not know the real distribution shape for each Kriging estimates.

Obviously, the generated local distributions could not cover the local conditional distribution space encountered in the Kriging. The problem may come from the linkage through the global distribution in the two spaces.

Conclusions

Although further work is required, we have shown an approach to simultaneously use DSS and reproduce the global distribution without ad-hoc post-processing or selective sampling. The procedure amounts to pre-calculate the conditional distribution shapes that will be needed. These shapes are calculated by back-transforming the theoretically correct shapes from Gaussian space using the theoretically correct back transformation procedure.

Acknowledgements

Chevron Petroleum Technology Company initiated this research and provided the seed funding to get it started. The funding supported Dr. YuLong Xie for six months of postdoctoral research at the Centre for Computational Geostatistics. We gratefully acknowledge Chevron's ideas, support, and openness with respect to publication.

Appendix: Program Description

The dssim.f90 program was created from the F90 version of sgsim. The global distribution used to establish the local distribution set could be the data itself or a reference distribution for the data. Local distributions are created by generating Gaussian distributions with varied mean (from –3.5 to 3.5) and variance (0 to 2, reason see later) and then transferred then into original data space.

The user specified the discretization levels for the mean, variance and number of quantiles. The local distributions are stored in memory for later retrieval in simulation.

The simulation follows the same procedure as that in other sequential simulation algorithms. Each node in the simulation domain is visited following a random path. Kriging mean and variance are obtained based on configuration of neighborhood. By comparing the Kriging mean/variance with those of the local distributions in the database, the distribution with the closest mean/variance from Kriging ones is selected for the simulation drawing purpose.

Since it is unlikely to find a distribution in the database with the exact mean/standard deviation values as the Kriging ones, the simulated value will be normalized to have the Kriging mean and variance.

$$v_{sim_K} = (v_{sim_F} - m_F) \times \frac{std_K}{std_F} + m_K$$

where v_{sim_F}, m_F, std_F are the simulated value, mean and standard deviation of the distribution from the database used for the simulation drawing and v_{sim_K}, m_K, std_K are the final simulation value, Kriging mean and standard deviation for the node.

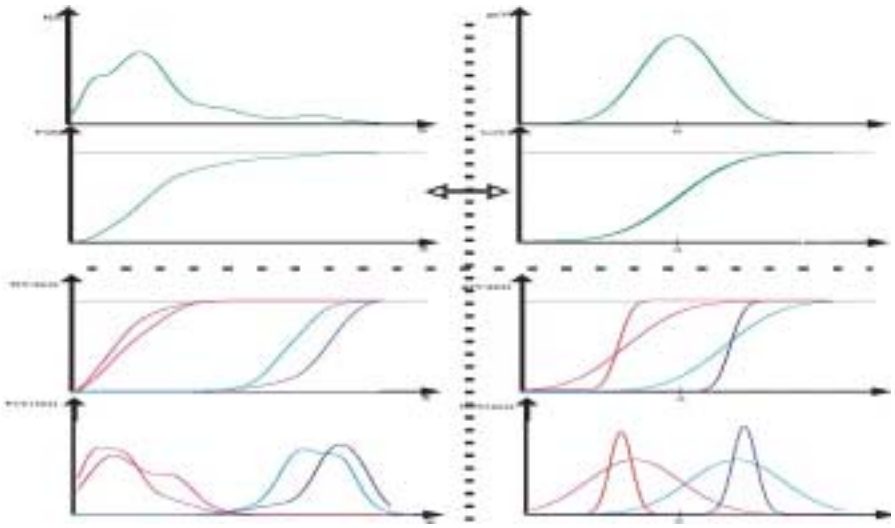

Figure 1: illustration of "real" *Z* space in the units of the data and Gaussian *Y* space where all conditional distributions are Gaussian. The top row shows the univariate *Z* and the univariate *Y* histogram. The second row shows the cumulative *Z* and *Y* distribution functions. This is the only way to go between *Z* and *Y* units. The lower two rows of figures shows histograms and cumulative distribution functions of conditional distributions (low and high mean / low and high variance).

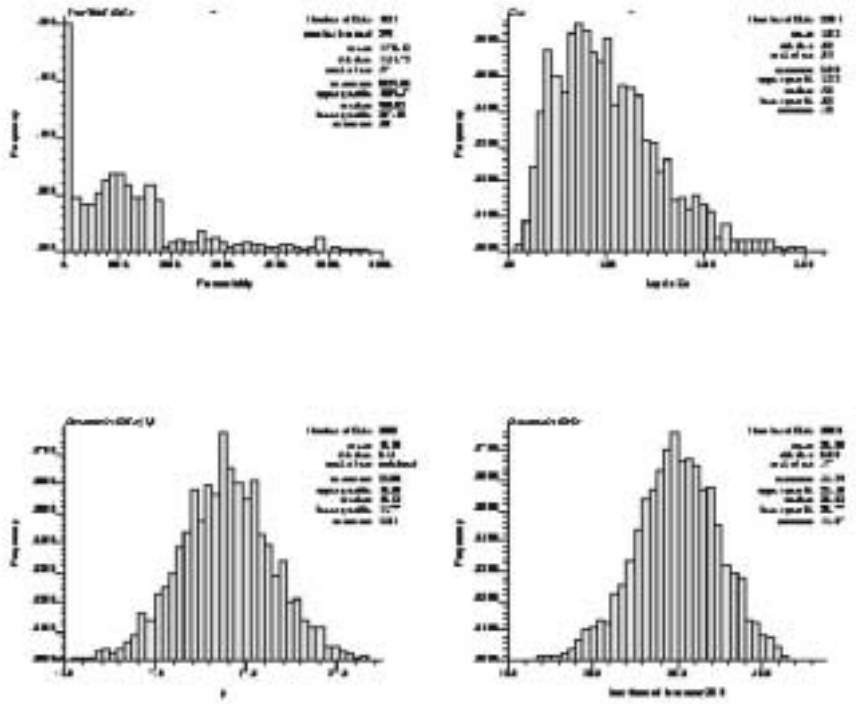

Figure 2: Histograms of the four data sets used to test the histogram reproduction capabilities of the proposed algorithm.

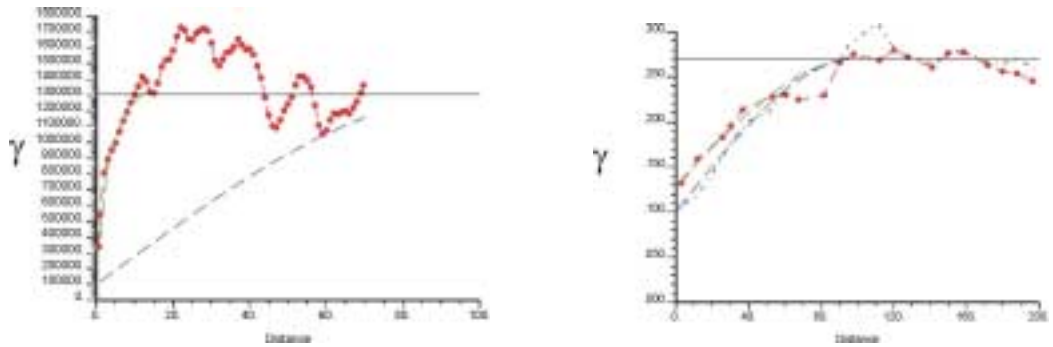

Figure 3: Variograms for the first two data sets. Note that these are not standardized since DSS requires the stationary variogram of the original Z variable.

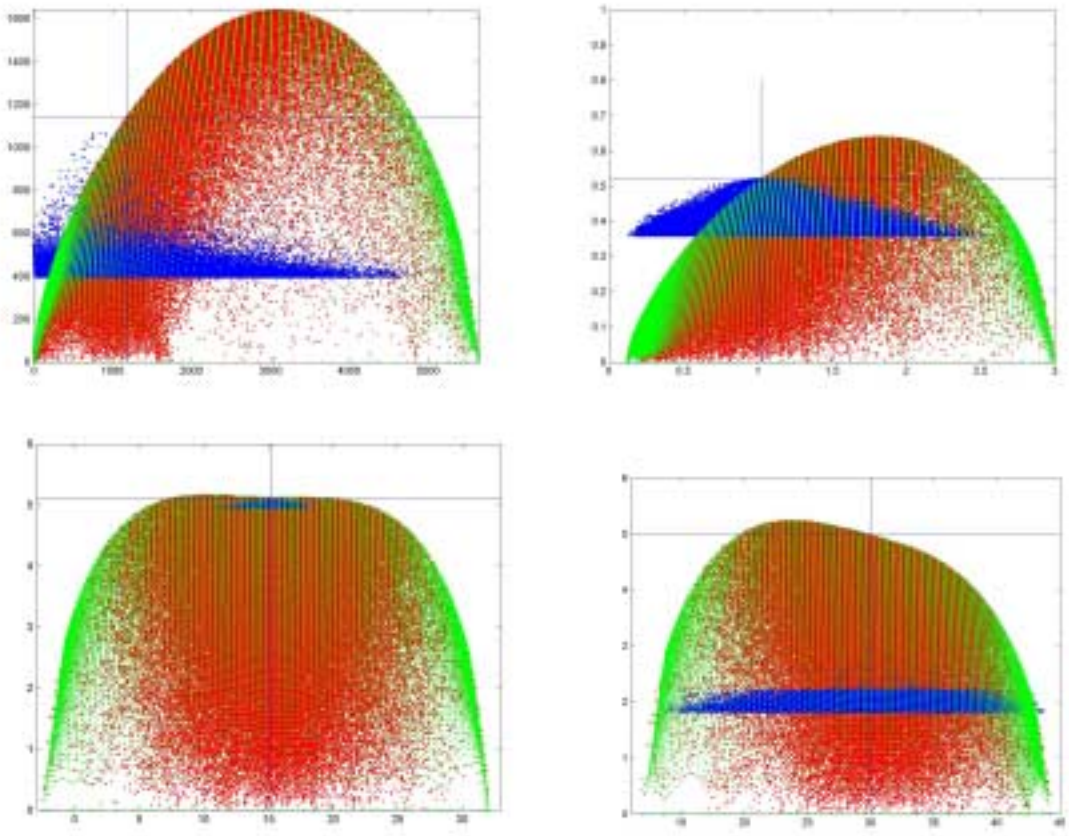

Figure 4: Domain of local distributions (red dots), approximation through discretization of 101 levels in mean and standard deviation (green dots), and the practical domain from Kriging (blue dots).

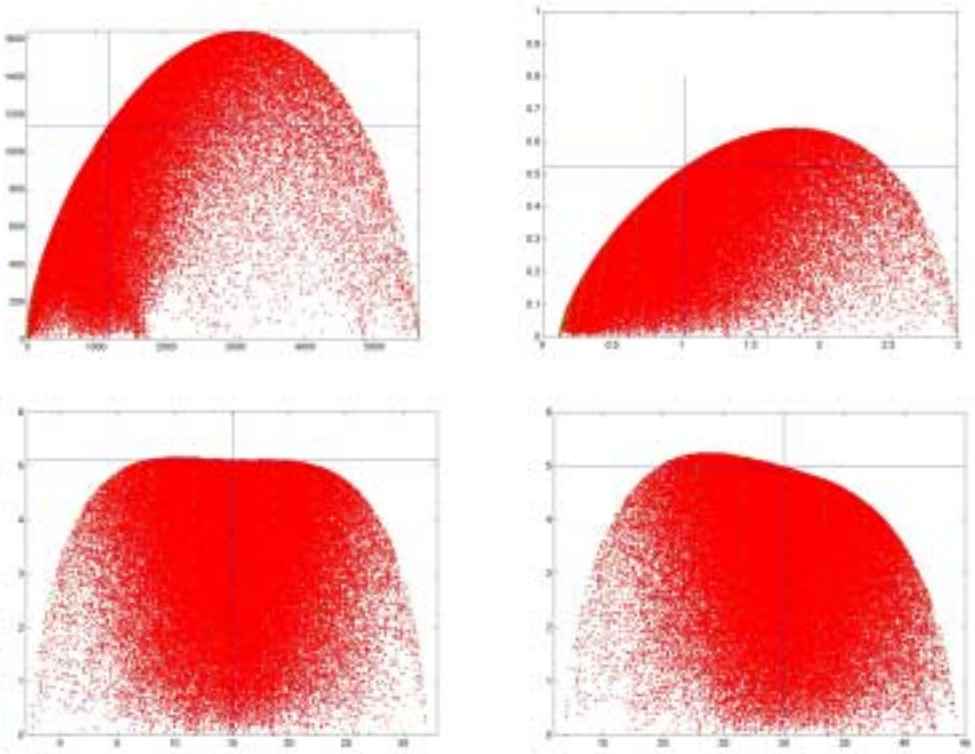

Figure 5: Domain of local distributions (red dots) and approximation through discretization of 11 levels in mean and standard deviation (green dots).

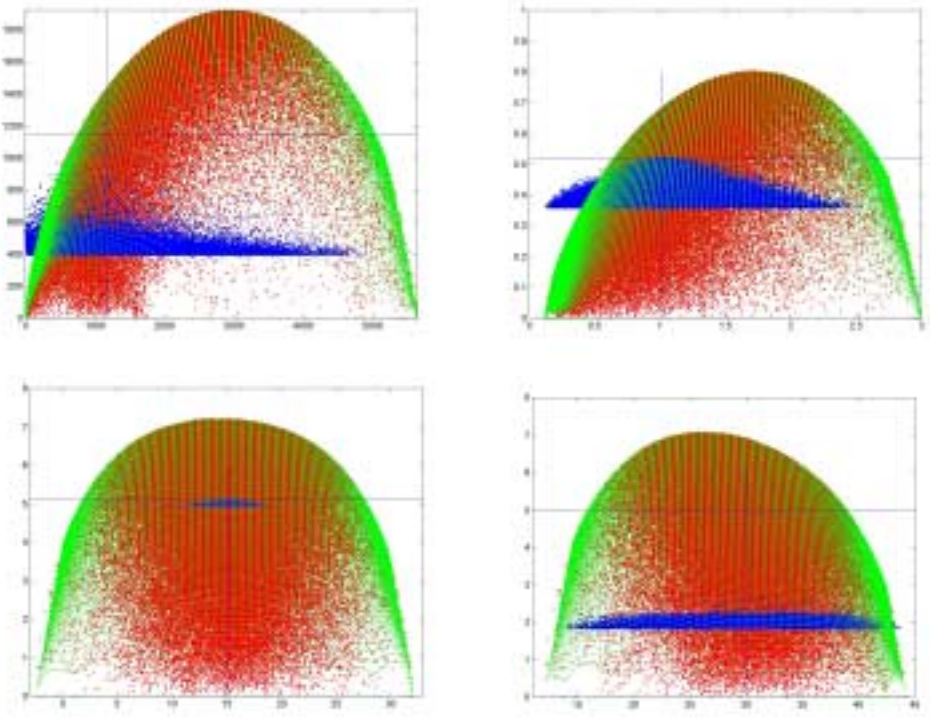

Figure 6. Domain of local distributions (red dots) and approximation through discretization of 101 levels in mean by increasing the variance in normal space from 1 to 2, and standard deviation (green dots).

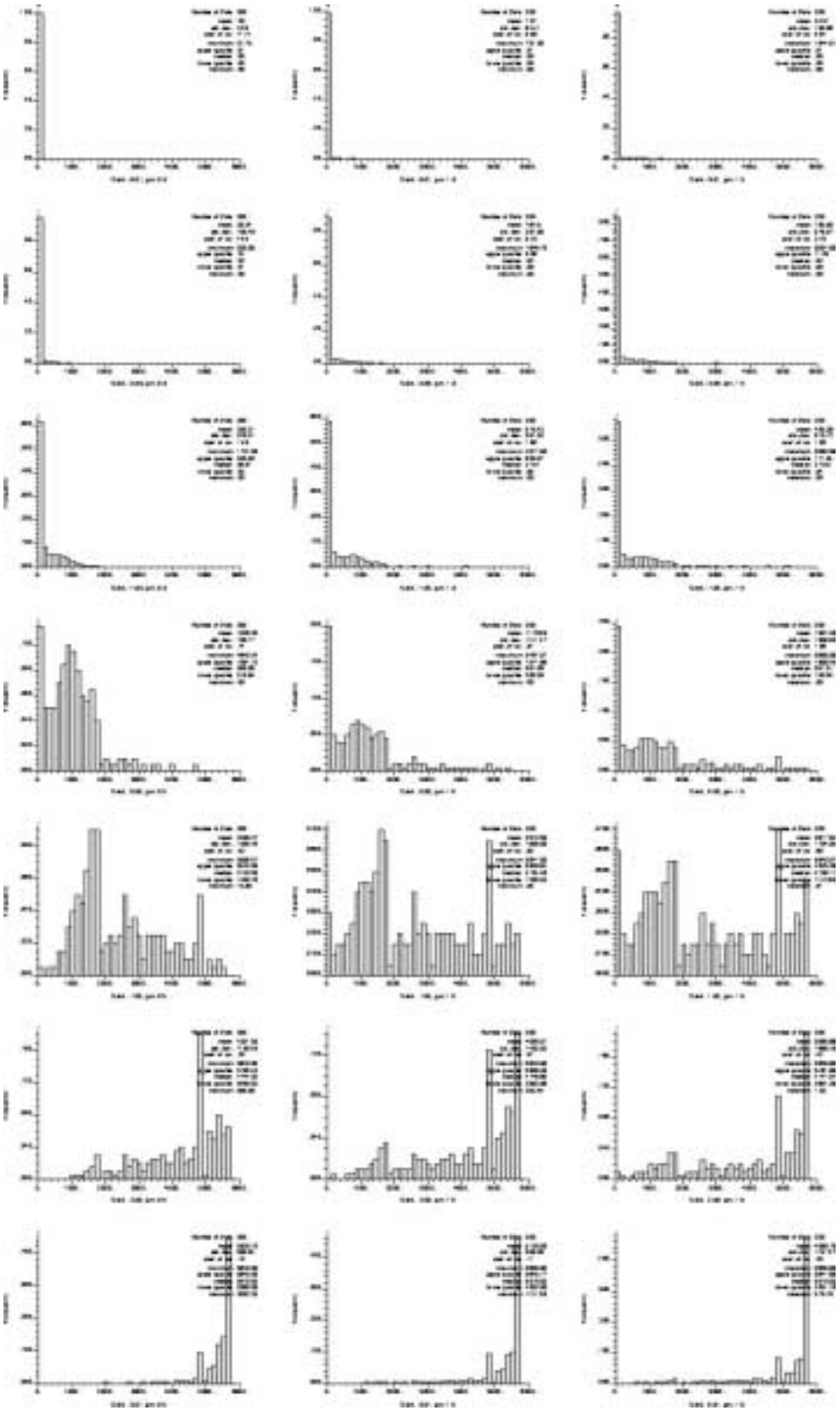

Figure 7: Local distributions of data set I (from left to right: variance in normal space: 0.5/1.0/1.5; from top to bottom: mean in normal space: -3.01/-2.03/-1.05/0/1.05/2.03/3.01)

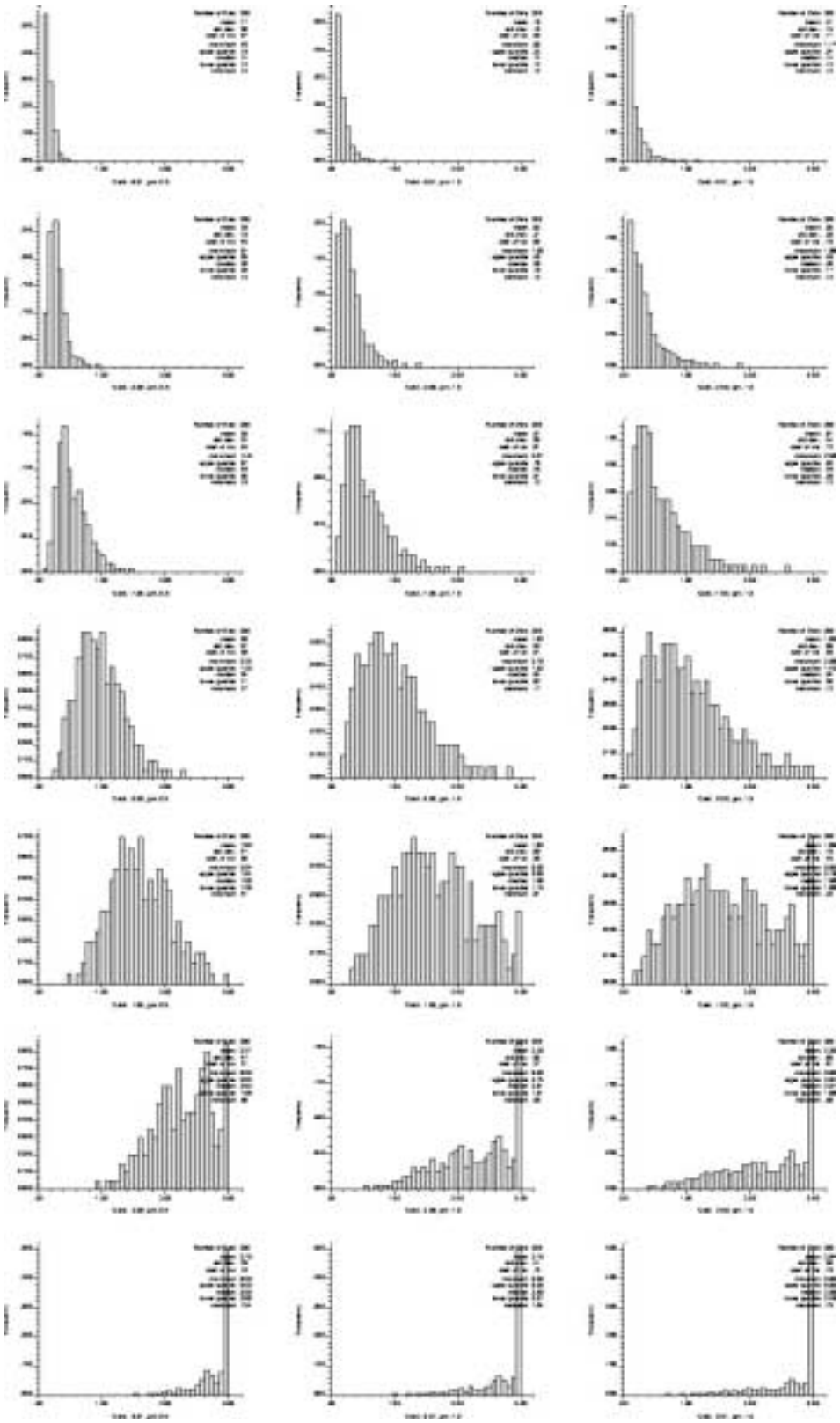

Figure 8. Local distributions of data set II (from left to right: variance in normal space: 0.5/1.0/1.5; from top to bottom: mean in normal space: -3.01/-2.03/-1.05/0/1.05/2.03/3.01)

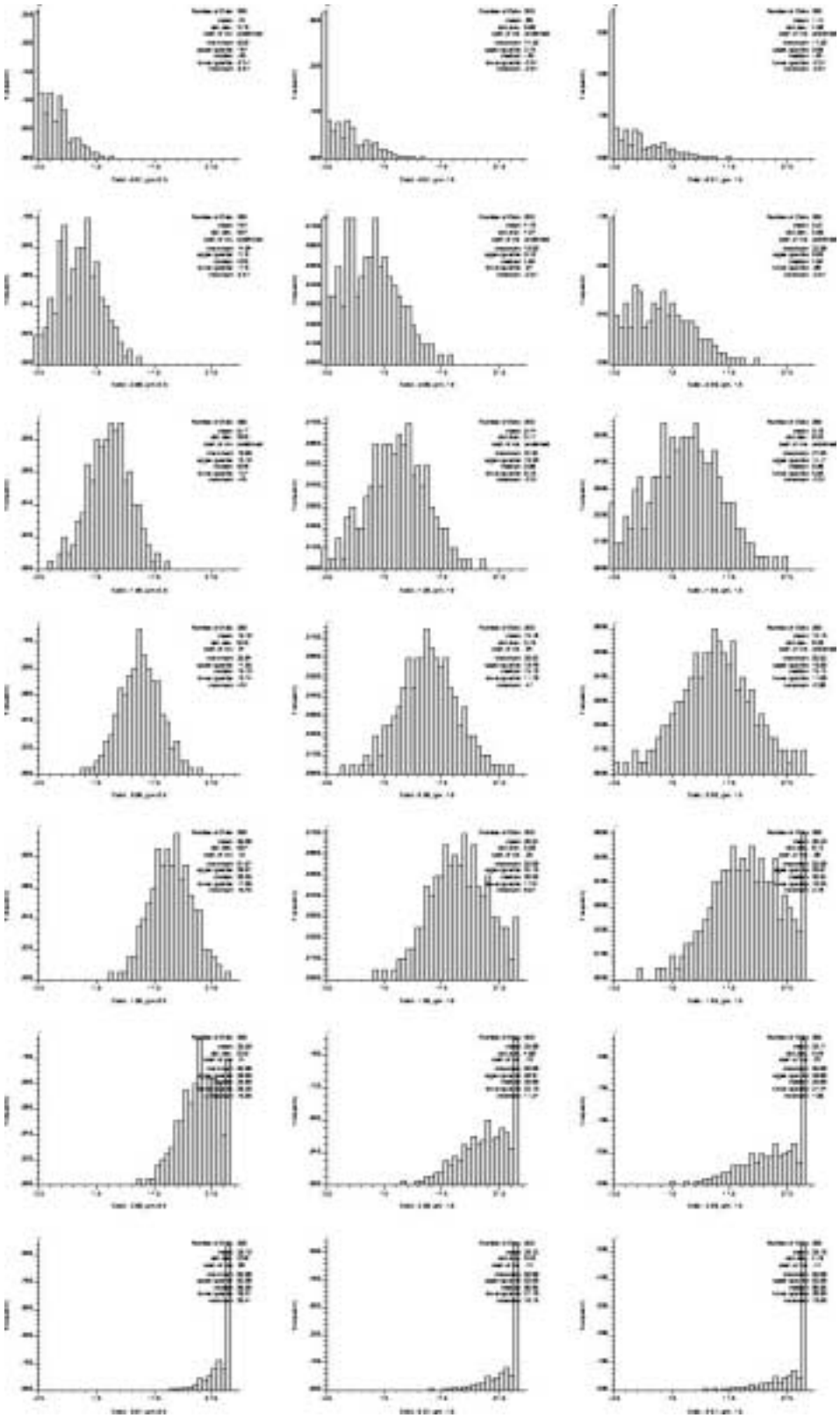

Figure 9: Local distributions of data set III (from left to right: variance in normal space: 0.5/1.0/1.5; from top to bottom: mean in normal space: -3.01/-2.03/-1.05/0/1.05/2.03/3.01)

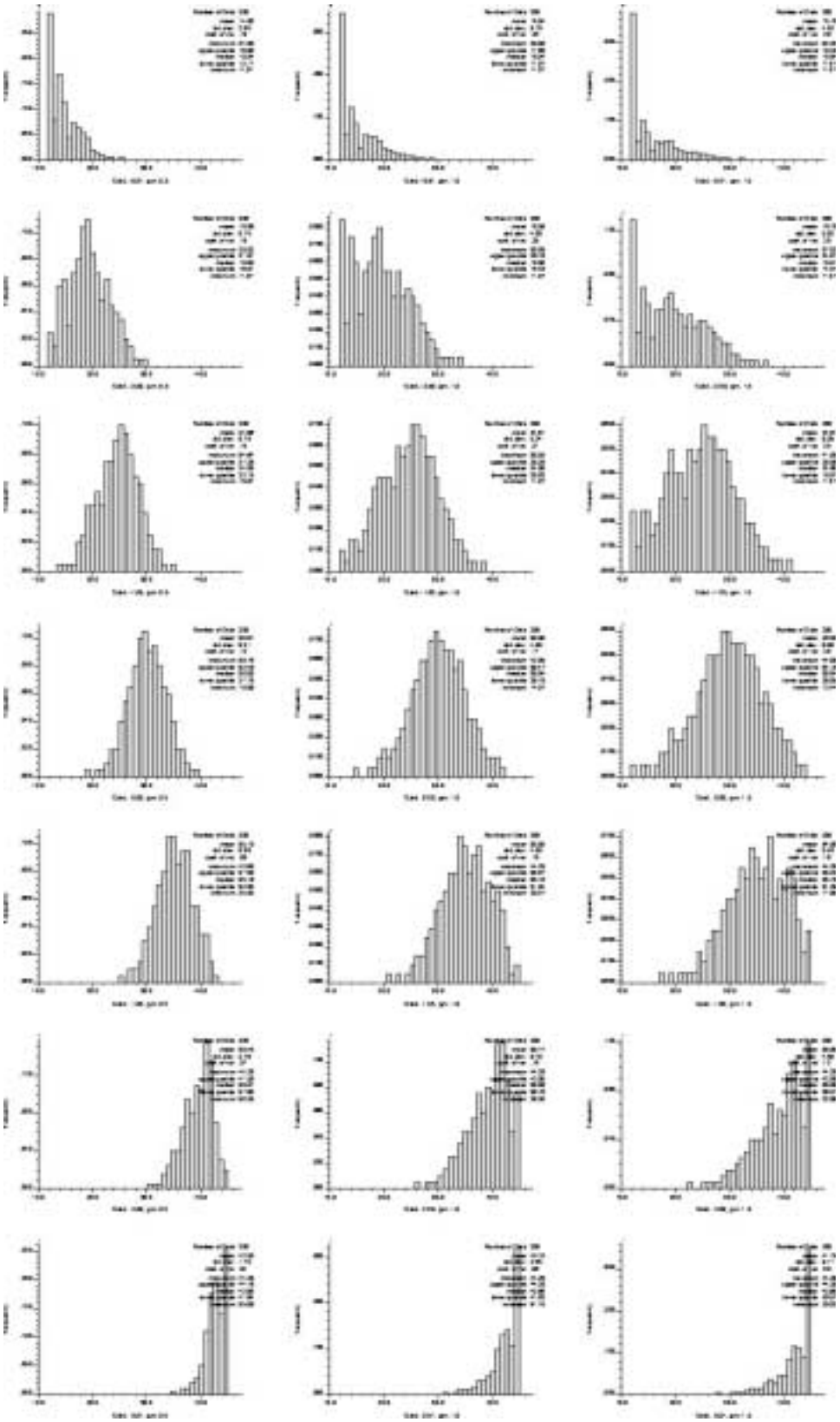

Figure 10: Local distributions of data set IV (from left to right: variance in normal space: 0.5/1.0/1.5; from top to bottom: mean in normal space: -3.01/-2.03/-1.05/0/1.05/2.03/3.01)

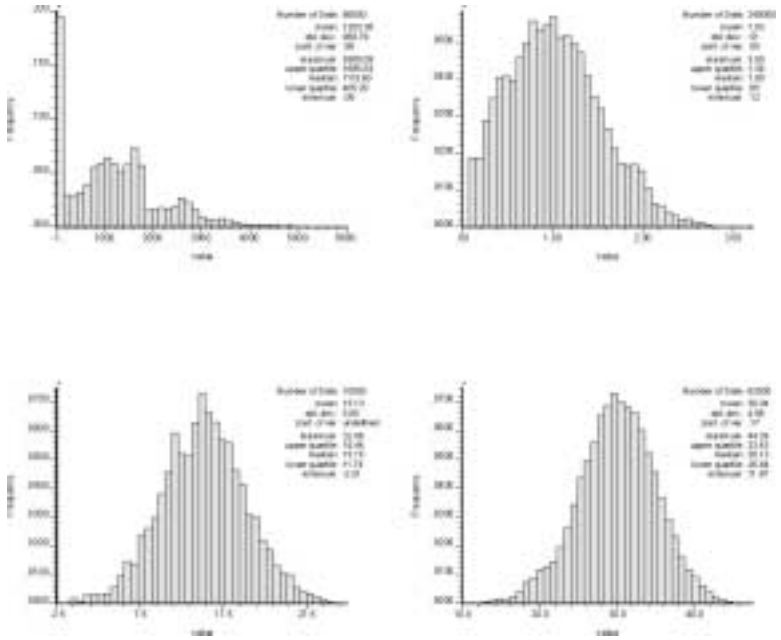

Figure 11. Reproduction of the histograms