

Computers & Geosciences 29 (2003) 39-51



www.elsevier.com/locate/cageo

DSSIM-HR: A FORTRAN 90 program for direct sequential simulation with histogram reproduction ☆

Bora Oz^a, Clayton V. Deutsch^{a,*}, Thomas. T. Tran^b, YuLong Xie^c

^a Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Canada T6G 2G7 ^b ChevronTexaco, Bakersfield, CA, USA ^c Pacific Northwest Labs, Richland, WA, USA

Received 24 October 2001; received in revised form 10 July 2002; accepted 12 July 2002

Abstract

Sequential simulation is a frequently used geostatistical simulation technique. The most widely used version of this technique is sequential Gaussian simulation, 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 since every conditional distribution is Gaussian with parameters given by kriging. Direct sequential simulation does not require any Gaussian assumption and simulates directly the data space; however, 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 lack of histogram reproduction is due to the unknown shape of the conditional distributions, which are used for drawing the simulated values.

We derive a simple and theoretically valid approach by establishing the shapes of the sequentially constructed conditional distributions. These shapes ensure histogram reproduction. The approach has been coded in FORTRAN 90 and called DSSIM-HR, where the extension HR refers to the feature of "Histogram Reproduction". © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Geostatistical simulation; Realizations; Multiscale data; Gaussian transformation

1. Introduction

Sequential simulation is often applied (Goméz-Hernández and Journel, 1993; Isaaks, 1990; Johnson, 1987). It 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 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 Gaussian space and the simulated y values are back transformed to original z data units. Secondary 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, 1997) is one implementation of the algorithm.

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

 $^{{}^{\}bigstar}\text{Code}$ on server at http://www.iamg.org/CGEditor/index.htm

^{*}Corresponding author. Tel.: +1-780-492-9916; fax: +1-780-492 0249.

E-mail addresses: boz@ualberta.ca (B. Oz), cdeutsch@ualberta.ca (C.V. Deutsch), jtttr@chevrontexaco.com (T.T. Tran), yulong.xie@pnl.gov (Y. Xie).

extreme values. Direct sequential simulation, presented later, does not effectively remove our reliance on the multivariate Gaussian distribution since the central limit theorem is acting in the kriging/averaging step of simulation. The main motivation for a "direct" method is based on the increasing need to simultaneously account for data at different scales, for example, core, well log, and seismic data in a petroleum reservoir modeling context.

Transformation of the data variable to a Gaussian distribution is problematic when dealing with data of different scale. The transform is non-linear and yet most averaging is linear (volumetric/mass proportions) or with very particular scaling laws (such as permeability). The non-linear transformation to a Gaussian variable entails that the Gaussian transformed value cannot be averaged linearly. The Gaussian transformation must be avoided to simultaneously account for multiscale data.

The notion of direct sequential simulation (DSS) was developed at the same time as sequential Gaussian simulation (Journel, personal communication, 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 each new simulated value is drawn randomly from local conditional distributions centered at the simple (co-) kriging estimates with a variance corresponding to the simple (co-)kriging variance. These distributions can be of any shape. Exercising this freedom, however, leads to simulated realizations where the univariate histogram is not controlled and therefore not reproduced. The histogram is important; it is a first-order statistic that has a first-order effect on the calculations made with the simulated realizations. The inability of DSS to reproduce the input histogram has been a significant problem. Aside from the Gaussian distribution, there is no general distribution shape that ensures global histogram reproduction (Caers, 2000a; Deutsch, personal communication, 2000).

A number of researchers have proposed the use of "post processing" in order to transform the resulting simulated values into another set of variables that reproduce the input global histogram (Journel and Xu, 1994; Caers, 2000a). This post processing is the same quantile-transformation procedure used to transform original Z variable to Gaussian Y variable. This procedure removes uncertainty (ergodic fluctuations) from the final histogram and, more importantly, it modifies any block data conditioning. Depending on the success of the original simulation at honouring the target distribution, severe adjustments may degrade the



honouring of variogram and may also introduce spatial discontinuities in the final result (Nowak and Srivastava, 1997).

Another approach has been to formulate an objective function as a measure of difference between the input global histogram and the histogram of the simulated values (Caers, 2000b). This objective function can be used to selectively accept/reject certain simulated values to ensure that the final realization reproduces the input global histogram. This approach can introduce artifacts near locations early in the simulation path and also removes ergodic fluctuations that are important for uncertainity analysis.

Nowak and Srivastava (1997) proposed that simulation proceed from a master list of values that exactly match the intended global distribution. Each value in the master list ends up being assigned to one grid node and is chosen by the results of simple kriging. At each step in the simulation the basic idea is to extract a subset of values from the remaining values whose weighted mean and variance are equal to SK results. The initial master list is depleted as simulation proceeds. There are some problems toward the end of each realization where there is no suitable subset.

Soares (2001) proposed an approach to reproduce the global histogram in DSS. The main idea of Soares's proposal is to draw the simulated values from those intervals of the global distribution that are centered at the simple kriging estimate. The interval range depends on the SK estimation variance. Two methodologies are proposed to define the intervals. The first approach is very similar to Nowak and Srivastava's idea of taking a subset of values that give the right mean and variance. The second approach uses the Gaussian transformation to determine the sampling intervals. The local SK estimate is transformed to a Gaussian value and the standardized SK variance are used to identify a Gaussian distribution to draw from; the drawn value is back transformed using the global distribution. Soares also extended this technique for the joint simulation of different variables. As mentioned in the original paper, there is a need for a local bias correction to account for the problem resulting from the non-linear transformation of the SK estimate to its normal score equivalent. This method appears similar to the one we implement; therefore, more discussion and a comparison with our proposal are given in an appendix.

Recently, Deutsch (2000) proposed an approach to reproduce the input global histogram in DSS. The key idea is to establish the shapes of the conditional distributions using the global normal-score or Gaussian transformation. There is a unique Gaussian distribution that can be back transformed to provide a valid *z*distribution with the correct mean and variance given by SK. The set of valid distributions can be calculated ahead of time using a look-up table. Gaussian transformation is used only to contruct the look-up table. The entire sequential simulation is performed with original data values. The overall histogram is effectively reproduced within statistical fluctuations without any aposterior transformation or correction schemes.

Tran et al. (2001) extended this proposal to account for multiple data sources. They showed that this proposal is successful in reconciling various data types including well, seismic and production data. DSS allows for more direct integration of seismic or production



Fig. 2. Three distributions used during application of DSSIM-HR: lognormal, bimodal and uniform.

(large scale) data whereas SGS only allows reproduction of the large scale data.

This paper documents and implements this last proposal. The new program is coded in FORTRAN 90. The code and the operation of the program are in the style of GSLIB (Deutsch and Journel, 1997). The approach and the implementation details are explained with illustrative examples.

2. Methodology

.300

.200

.100

.000

.0

5.0

10.0 15.0

data

Frequency

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 data scale. The classical steps of sequential simulation:

- 1. Randomly choose a location **u** to be simulated.
- 2. Calculate the kriging estimate and variance using the original data and all previously simulated values.

mean 1.74

std. dev. 1.26

20.0 25.0 .160

.080

.040

.000

.0

5.0

10.0 15.0

data

Frequency .120

- 3. Draw a value from a distribution with a mean equal to the kriging estimate and a variance equal to the kriging variance.
- 4. Return to step 1 until all nodes have been simulated.

The multiGaussian approach is typically used (SGS) 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 global histogram, however, will typically depend on the shape of the local distributions, the data distribution, 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

mean 6.32

std. dev. 4.29

.100

080

.060

.040

.020

.000

.0

5.0

10.0

15.0 20.0 25.0

data

Frequency

mean 3.33

std. dev. 2.39

20.0 25.0



Fig. 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).

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.

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

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

$$z = F_Z^{-1}(G(y)).$$
 (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 (data) space can be determined from back transformation of non-standard univariate Gaussian distributions by Monte–Carlo simulation or by back transformation of L regularly spaced quantiles, p^l , l = 1, ..., L:

$$z^{l} = F_{Z}^{-1}[G(G^{-1}(p^{l})\sigma_{y} + y^{*})], \quad l = 1, ..., L,$$
(3)

where y^* and σ_y are the mean and standard deviation of the non-standard Gaussian distribution, and the p^l , l = 1, ..., L are uniformly distributed values between 0 and 1.



Fig. 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).

The distribution of uncertainty in Z space, denoted $F_{Z,y^*,\sigma_y}(z)$, is assembled from the z^l , l = 1, ..., L values. The Gaussian parameters, y^* and σ_y , are added as subscripts to the z-distribution to identify where it came from. Although these distributions retain some of the characteristics of the global distribution, $F_Z(z)$, the shape of the conditional distributions, $F_{Z,y^*,\sigma_y}(z)$, are neither the original Z data distribution nor a Gaussian distribution.

The predetermined distribution shapes can be used in the direct sequential simulation algorithm. All kriging and simulation is performed in original Z variable units. The Gaussian transform is only used for obtaining 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. Look-up table construction: Generate non-standard Gaussian distributions by choosing regularly spaced mean values (approximately from -3.5 to 3.5) and variance values (approximately from 0 to 2). Then,

using Eq. (3) calculate and store the z-conditional distributions $(F_{Z,y^*,\sigma_y}(z))$, and their mean and variance values.

- 2. Calculate the mean and variance of the local distribution, $z^*(u)$ and $\sigma_z^2(u)$ in original z units by simple (co-)kriging using all relevant original data and previously simulated values.
- 3. Retrieve the closest z-conditional distribution, $F_{z,y^*,\sigma_y}(z)$, from the look-up table by searching for the one with the closest mean and variance to the (co-) kriging values.
- 4. A simulated value is drawn from the *z*-conditional distribution by Monte–Carlo simulation, that is, $z^s = F_{z,y^*,\sigma_y}^{-1}(p)$ where z^s is the simulated value and *p* is a random number uniformly distributed between 0 and 1, U(0, 1).

This approach will create realizations that reproduce the (1) local point and block data in the original data units, (2) the mean, variance, and variogram of the Z variable, and (3) the histogram of the Z variable.



Fig. 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).

The user specifies the discretization levels for the Gaussian mean, variance, and number of quantiles for the lookup table. 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 via a fast searching algorithm. The closest distribution will not have the exact mean and variance; therefore, we rescale slightly the closest distribution to have exactly the correct mean and variance.

The required variogram is calculated from the original data with no normal or Gaussian transformation. All statistical input comes from original data units.

3. Program details

The DSSIM-HR program is based on the FORTRAN 90 version of the SGSIM program from GSLIB. The

mean 3.47

following presents the details specific to the implementation of DSSIM-HR.

The bldcond module builds the lookup table of conditional distribution shapes. The input to this module includes (1) the available data, (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 calculated in original units.

The kriging module returns with the local estimate, K_{est} , and local estimation variance, K_{std} . The getcond module selects the closest local distribution from the database.

The drawcond module draws a simulated value by Monte–Carlo simulation from the nq quantiles. The simulated value is rescaled so that it is drawn from a distribution with the correct mean and variance:

$$v_{sim_{node}} = (v_{sim} - m_{sim}) \cdot \frac{K_{std}}{std_{sim}} + K_{est},$$
(4)

mean 3.49



Fig. 6. Reproduction of histograms for four different realizations generated by DSSIM-HR using input lognormal distribution.

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 to the next grid node.

3.1. New parameters

The parameters are almost exactly the same as the SGSIM program. An example parameter file is given in Fig. 1. The z-local distributions of uncertainty and their means and variances must be generated. These files are created during the first execution of the DSSIM-HR program; there is an option in the program so that they do not have to be created again for subsequent simulation runs.

The user specifies the discretization levels for mean, variance, and the number of quantiles. For the parameter file given in Fig. 1, the input values are 170, 170,

and 300. Such detailed discretization will result in excellent precision in reproduction of the input global distribution.

An important advantage of DSSIM-HR is that the simulation is performed 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.

4. Example applications

A lognormal, bimodal and uniform distribution are used in the examples shown below; see Fig. 2. Arbitrary 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 are significantly different from either the widely assumed Gaussian distribution or the original global histograms. Consider



Fig. 7. Reproduction of histograms for four different realizations generated by DSSIM-HR using input bimodal distribution.

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 distributions are defined by Eq. (3). The distributions for the lognormal distribution are shown in Fig. 3. The boxed histogram corresponds to a Gaussian mean and variance of 0.0 and 1.0, yielding the original distribution.

Some selected distributions for the bimodal and uniform distributions are shown in Figs. 4 and 5. The shapes of distributions differ from the global distributions and the Gaussian distribution. The skewness of the distributions changes significantly. A recent study by Oz (personal communication, 2002) showed that Hermite polynomials and Disjunctive kriging would give the exact same conditional distribution shapes.

4.1. Histogram reproduction

A 2-D simulation field of 150 by 150 grid nodes was considered with an exponential variogram model with 20% nugget effect. Other variograms were also considered and the variograms are reproduced in all cases. Four realizations were generated and the resulting histograms are shown in Figs. 6–8. In all the cases the realizations successfully reproduce the global histogram and summary statistics such as the mean and variance. There are statistical (ergodic) fluctuations that are due to a finite domain size.

4.2. Variogram reproduction

Variogram reproduction is theoretically guaranteed; 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 Fig. 9. Note the excellent reproduction in all cases.

5. Application to porosity data

Porosity data from 44 wells over a 5 km by 5 km area were considered. The location map of the field is given in



Fig. 8. Reproduction of histograms for four different realizations generated by DSSIM-HR using input uniform distribution.



Fig. 9. Reproduction of variogram for lognormal, bimodal, and uniform distribution.

Fig. 10 (top figure). The histogram and the cumulative distribution are also shown in Fig. 10. The ominidirectional experimental variogram was calculated and fitted with an isotropic exponential variogram model with a 1.3 km range.

The study area was discretized with 200 by 200 grid nodes of size $\Delta x = \Delta y = 250$ m. The look-up table for local conditional distributions was generated. One simulated realization (bottom left figure in Fig. 10) for the porosity was generated. The histograms on the third row of Fig. 10 show close reproduction of the input global histogram. The bottom right figure shows the excellent variogram reproduction; the line is the input model and the circles are calculated from the simulated realization.

6. Conclusions

In traditional SGS, a multivariate Gaussian assumption is taken after univariate transformation to a Gaussian histogram. The value of working in original data units, instead of transforming to a Gaussian histogram, permits straightforward integration of multiscale data. Mean, variance, and variogram reproduction is guaranteed with all implementations of sequential simulation. Previous efforts to work in "direct" data units have had 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 were shown to demonstrate the successful histogram reproduction capability of DSSIM-HR.

Appendix

Soares (2001) proposed a solution methodology similar to our proposal; he also proposed to use Gaussian transformation to get the shapes of the conditional distributions. The main idea of Soares's proposal is to sample from the global distribution using the local simple (co-)kriging estimate and variance. The Gaussian transformation is used to determine the sampling interval. The normal score transformed local SK estimate and with the SK standardized estimation variance are used to identify the interval of the global distribution to be sampled. The aim of this appendix is to highlight the similarities and differences between Soares's approach and the one implemented.

Both proposals establish parameters of z-conditional cumulative distributions in direct space using SK. This gives z^* and σ_{KZ}^2 . Both proposals also calculate



Fig. 10. Top: Location map for porosity data, Second row: Original input histogram and cdf, Third row: Simulated histogram and cdf, Fourth row: One sample simulated realization and variogram reproduction (line for variogram model and circles are from simulated realization).



Lognormal Distribution

Fig. 11. Scatterplot of $Y^*_{DSSIM-HR}$ vs. Y^*_{Soares} (blue squares) for lognormal distribution. Pink line is 45° line stands for situation of $Y^*_{Soares} = Y^*_{DSSIM-HR}$.



Fig. 12. Scatterplot of $Y^*_{DSSIM-HR}$ vs. Y^*_{Soares} (blue squares) for bimodal distribution from two different 2D field simulation. Pink line is the 45° line stands for situation of $Y^*_{Soares} = Y^*_{DSSIM-HR}$.

non-standard Gaussian parameters y^* and $\sigma_{K,Y}^2$. The major difference is in how these parameters are estimated; see next. Both proposals draw from the non-standard Gaussian cdf, $y^s = G^{-1}(p)\sigma_{K,Y} + y^*$, and back transform to direct space: $z^s = F_Z^{-1}(G((Y^s)))$.

In our proposal, we determine the y^* and $\sigma_{K,Y}^2$ that *exactly* reproduce the SK estimate z^* and variance $\sigma_{K,Z}^2$. In Soares's proposal, y^* and $\sigma_{K,Y}^2$ are calculated as

$$v^* = G^{-1}F(z^*)$$

and

$$\sigma_{K,Y}^2 = \sigma_{K,Z}^2 / \sigma^2,$$

where σ^2 is the variance of the input global histogram.

Due to non-linear characteristics of the normal score transform $E\{y^s\} = y^*$ but $E\{z^s\} \neq z^*$. The Gaussian distribution is not centered at the correct value and does not have the correct variance. This leads to the bias correction that should be done for each simulated grid block during the direct sequential simulation. The magnitude of each correction depends on the difference between SK estimate z^* and the mean value that the observed value after back transformation of the approximate Gaussian parameters. This bias correction is similar to the variance term in lognormal kriging.

We simulated a 100 by 100 2D field with a reference lognormal distribution and plotted the $y^*_{DSSIM-HR}$ vs. y^*_{Soares} . As we can see from Fig. 11, most of the values do not fall on the 45° line. Next, we used a bimodal distribution; the corresponding cross-plots ($y^*_{DSSIM-HR}$ vs. y^*_{Soares}) are shown in Fig. 12. There is considerable discrepancy that must be corrected somehow.

References

- Caers, J., 2000a. Adding local accuracy to direct sequential simulation. Mathematical Geology 32 (7), 815–850.
- Caers, J., 2000b. Direct sequential indicator simulation. In: Kleingeld, W.J., Krige, D.G. (Eds.), Proceedings of the Geostatistics 2000. Cape Town, South Africa, pp. 39–48.
- Deutsch, C.V., Journel, A.G., 1997. GSLIB: Geostatistical Software Library and User's Guide, 2nd Edition. Oxford University Press, New York, 339pp.
- Goméz-Hernández, J., Journel, A.G., 1993. Joint sequential simulation of multiGaussian fields. In: Soares, A. (Ed.), Geostatistics Troia 1992, Vol. 1. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 85–94.
- Isaaks, E.H., 1990. The application of Monte Carlo method to the analysis of spatially correlated data. Ph.D. Dissertation, Stanford University, Stanford, CA, 213pp.
- Johnson, M., 1987. Multivariate Statistical Simulation. Wiley, New York, 237pp.
- Journel, A.G., Xu, W., 1994. Posterior identification of histograms conditional to local data. Mathematical Geology 26 (6), 323–359.

- Nowak, M.S., Srivastava, R.M., 1997. A geological conditional simulation algorithm that exactly honours a predefined grade-tonnage curve. In: Baafi, E.Y., Schofield, N.A. (Eds.), Proceedings of the Geostatistics Wollongong 96, Vol. 2. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 669–682.
- Soares, A., 2001. Direct sequential simulation and cosimulation. Mathematical Geology 33 (8), 911–926.
- Tran, T.T., Deutsch, C.V., Xie, Y., 2001. Direct geostatistical simulation with multiscale well, seismic, and production Data. SPE Annual Technical Conference and Exhibition, New Orleans, September 30–October 3, SPE Paper Number 71323.